Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.16, No.1, 2025

The Use of Multiple-choice Questions as an Assessment Tool in First-year University Physics Modules

N. Bhaw* J. Kriek G.J. Rampho

Department of Physics, University of South Africa, PO box 392, 0003, South Africa

* E-mail of the corresponding author: bhawn@unisa.ac.za

Abstract

The effectiveness of multiple-choice questions (MCQs) as an assessment tool has been a subject of interest in educational research. This study investigates the effectiveness of MCQs at the item level, focusing on aspects such as the difficulty index, discrimination index, distractor effectiveness, and overall reliability of the MCQ assessment. The research questions aim to provide insights into the performance and quality of MCQ items used in first-year modules over two years. Findings indicate that although the assessments are, on average, acceptably difficult, certain questions are too difficult. This difficulty is due to inappropriately designed MCQs, which is evidenced by the low overall reliability of the assessments. There is a statistically significant strong positive correlation between conformity to MCQ design guidelines and average assessment scores. The findings encourage ongoing discourse on the use of MCQs as an assessment tool to inform educators and policymakers about the strengths and weaknesses of MCQ design .

Keywords: multiple-choice questions, difficulty index, discrimination index, distractor index, reliability, multiple-choice question guidelines

DOI: 10.7176/JEP/16-1-07
Publication date: January 30th 2025

Introduction

Multiple-choice questions (MCQs) are widely used in science education due to their efficiency in measuring students' knowledge across various disciplines (Siddiqui, 2022). MCQs offer versatility, objectivity, and the ability to assess various content areas (Anunpattana et al., 2023). They can cover diverse scientific topics and evaluate factual knowledge, conceptual understanding, problem-solving skills, and critical thinking abilities (Sideris et al., 2022). The objective scoring of MCQs minimizes subjectivity, making them suitable for large-scale assessments and standardized tests (Baldwin et al., 2022). Well-designed MCQs promote critical thinking by requiring students to apply knowledge, solve complex problems, and evaluate explanations (Alkhatib, 2022). However, MCQs may not capture all aspects of scientific abilities, such as hands-on skills or creativity (Fadzil et al., 2022). Effective MCQs depend on design and contextual factors, including administration, conditions, timing, instructions, and format. The present study focuses on MCQ design effectiveness, measured by the difficulty index (DF), discrimination index (DI), distractor effectiveness (DE), reliability (KR-20), and MCQ conformity to design guidelines (CDG).

Research Questions

The research questions guiding the present study are:

How do the MCQ items' difficulty, discrimination between high and low achievers, suitability of distractors, and overall reliability contribute to the effectiveness of MCQ-based summative assessments?
How do the calculated MCQ indices reflect student performance in relation to their conformity with design guidelines and assessment scores?
What are the main characteristics of poorly designed MCQs?

The research questions aimed to provide insights into the performance and quality of MCQ items used in first-year Physics modules at an open and distance learning university over a two-year period. Analyzing the DF, DI, DE, and KR-20 provides valuable insights into the individual items and the overall test, enabling educators to enhance MCQ assessments' validity, reliability, and fairness (Anunpattana et al., 2023).

Literature Review

MCQs have been a long-standing assessment tool (Coombs et al., 1956) with a history of extensive use and are prevalent in various educational settings due to their advantages, including efficient administration, ease of scoring, and the ability to assess a wide range of knowledge and skills (Nojomi & Mahmoudi, 2022). However, researchers have continually explored the effectiveness and reliability of MCQs as an assessment method. Early studies on MCQs in assessment primarily aimed to understand their validity, reliability, and impact on student performance.

For instance, Tatsuoka (1983) analyzed the relationship between item difficulty and discrimination to assess the effectiveness of MCQs. Sijtsma and Molenaar (1987) investigated the reliability of MCQ assessments using methods such as item response theory. These early studies formed the foundation for subsequent MCQ effectiveness and reliability research. Measuring the quality and characteristics of individual MCQ items is a crucial aspect of assessing their effectiveness (Hassen, 2022). Standard measures include the difficulty index (DF), discrimination index (DI), and distractor effectiveness (DE). The DF reflects "the proportion of students who answer an item correctly, while the discrimination index indicates the item's ability to differentiate between high- and low-performing students" (Popham, 1990, p. 220). Researchers such as Tollefson (1987) found that there is a significant relationship between DF and DI in MCQ items, indicating that well-designed MCQs can effectively differentiate between high and low achievers. Considine et al. (2005) discovered that the design, format, validity and reliability of MCQs are crucial for accurate assessment. These findings are particularly relevant to the present study, as they highlight the importance of carefully designing MCQs to ensure they are challenging yet fair, and capable of reliably assessing students' understanding and performance.

Reliability, which refers to the consistency and stability of scores obtained from an assessment, is vital in MCQ assessments (Peeters et al., 2021). Classical Test Theory (CTT) measures, such as Cronbach's alpha and Kuder-Richardson formulas, traditionally estimate the internal consistency of MCQ assessments (Kumar et al., 2021). Ang and Boo (2006) explored the impact of MCQ usage and the development of students' thinking skills using reflective MCQ as a tool in formative assessment. Recent years have witnessed advancements in methodologies and technologies that have furthered MCQ effectiveness and reliability research (Jiang et al., 2022). Computerized adaptive testing (CAT) has emerged as an innovative approach to enhance MCQ assessments by dynamically adjusting item difficulty based on individual student responses (Ijiwade & Alonzo, 2023).

Studies by Mittelhaëuser et al. (2015) and Xu et al. (2020) explored the application of CAT in MCQ assessments, providing insights into its potential benefits and challenges. Moreover, the use of Item Response Theory (IRT) models has gained prominence in assessing the psychometric properties of MCQ items and improving assessment reliability (Kumar et al., 2023). Smith et al. (2020) utilized IRT models to investigate MCQ items' difficulty and discrimination parameters, enhancing understanding of item functioning. Al-zboon (2022) studied the impact of the number of distractors in MCQs and concluded that the four and five-distractor structure had higher values of the test items' information function and discrimination parameters compared to the three-distractors structure. The key studies reviewed mostly focus on measuring the effectiveness and reliability of MCQs in educational assessments. These studies highlight the significance of assessing MCQ effectiveness using various metrics such as the DF, DI, and DE. By examining these metrics, researchers have gained insights into student performance and the quality of MCQ assessments, shedding light on their efficacy and reliability for measuring student knowledge and understanding. Ahmad (2019), Chit (2020), Sharma (2021), and Uddin et al. (2022) contribute to the understanding of MCQ effectiveness by examining the difficulty index. Sharma emphasizes the importance of DI, and Reza et al. (2021) investigated DE. Additionally, Uddin et al. and Iqbal et al. (2023) provide valuable insights into the overall measurement of MCQ effectiveness and reliability. Collectively, these studies enhanced the understanding of how to evaluate the efficacy and reliability of MCQ assessments using key metrics, which can inform future assessment practices and improve the quality of educational evaluations. The key studies provided a framework to interpret and discuss the current study's findings.

Conceptual Framework

A robust framework for MCQ analysis ensures the quality and effectiveness of MCQ assessments (Adnan et al., 2023). The conceptual framework allowed for the interpretation of quantitative and qualitative results. Interpretation of the quantitative results was based on IRT, which was first proposed during the 19th century in the field of mathematics and psychology. Researchers such as Lawley (1943) and Lord (1956) propelled the theory's advancements over the past 60 years. Research by Lazarsfeld (1950), Rasch (1960), Wright (1968) and Andrich (1978) also significantly contributed to IRT's development and widespread application in educational measurement. The quantitative measurements, based on IRT, include the DF, DI, DE, and overall test reliability. The qualitative analysis is framed on the guidelines for writing selected response items (Haladyna et al., 2019).

Materials and Methods

Sample

The present study analyzed three first-year university physics theory modules that form part of the BSc Physics program. The modules analyzed were Mechanics (X), Electromagnetism (Y), and Modern Physics. (Z), evaluated over two years, Year A and Year B. The individual assessments were labelled XA, XB, YA, YB, ZA, and ZB.

Preparation and Coding of Data

The analysis of each assessment produced a response table listing items in rows and respondents in columns for each module. The correct option of an MCQ is defined as the key, and the incorrect options are defined as distractors. Responses matching the key were coded as 1, otherwise as 0. The response table includes a total score for each respondent, calculated as the sum of coded responses, and is ordered from highest to lowest total. Students in the top quartile form the high-achievers group, while those in the bottom quartile form the low-achievers group. Nonachievers are students who respond incorrectly to the MCQ item. The correct response rate for each option is calculated as a ratio of the number of correct responses to the number of responses. Distractors with a correct-response rate of less than 5% are defined as non-functional distractors (NFDs).

Data Analysis Tools

A special case of Cronbach's test, known as the Kuder-Richardson Formula 20 (KR-20; Iqbal et al., 2023) was calculated as the reliability of each assessment item. The analysis at an item level, calculated the DF, DI, DE, KR-20, and CDG.

Difficulty Index (DF)

The DF measures the ease or difficulty of an individual MCQ item by calculating the percentage of students who answered it correctly (Ansari et al., 2022; Iqbal et al., 2023; Sharma, 2021). By analyzing the DF, educators can align the difficulty of MCQ items with desired learning outcomes and students' abilities, identifying problematic items for revision or elimination. The DF was calculated by substituting the number of correct responses from high achievers (H), low achievers (L), and the combined total number of respondents (N), into Equation 1.

$$DF = \left(\frac{H \mid L}{N}\right) \times 100 \tag{1}$$

The literature reviewed indicates that an ideal DF should be between 50% and 60% for a valid MCQ (Ansari et al., 2022; Iqbal et al., 2023; Sharma, 2021). The criteria for the present study rate a DF of less than 30% as too difficult; between 30% and 70% as acceptable difficulty, and greater than 70% as too easy.

Discrimination Index (DI)

The DI assesses an MCQ item's ability to differentiate between high-achieving and low-achieving students (Iqbal et al., 2023). Analyzing the DI helps educators identify biased or ineffective items and ensures that MCQs contribute to valid and reliable assessments of students' knowledge and skills (Nojomi & Mahmoudi, 2022). The DI was calculated by substituting the number of correct responses from high achievers (H), low achievers (L), and the combined total number of respondents (N) into Equation 2.

$$D = \left(\frac{2 \times (H - L}{N}\right) \tag{2}$$

Based on Iqbal et al. (2023), the criteria for the present study rates a DI less than or equal to 20% regarded as poor, between 21% and 24% as acceptable, between 25% and 34% as good, and a DI greater or equal to 35% as excellent.

Distractor Effectiveness (DE)

The DE measures the quality of an MCQ item's incorrect options (distractors). Effective distractors should be plausible and appealing to students who lack the necessary knowledge while being unappealing to knowledgeable students (Iqbal et al., 2023; Jia et al., 2020; Kumar et al., 2023). Effective distractors are chosen by many students, indicating confusion, while ineffective distractors are rarely selected, failing to differentiate between varying levels of student knowledge (Kucwaj et al., 2022). Designing assessment items with effective distractors enables educators to replace learner misconceptions (Tolba & Youssef, 2024). Evaluating DE allows educators to refine item options, eliminate ineffective distractors, and improve the item's ability to discriminate between different levels of understanding (Mendez-Carbajo, 2023). The DE was calculated by substituting the number of total distractors (D) and the number of non-functioning distractors (NFDs) into Equation 3.

$$DE = \left(\frac{D - NFD}{D}\right) \times 100 \tag{3}$$

An MCQ, having five options and four NFDs, results in a DE of 0, which is regarded as unacceptable. If there are three NFDs, the DE is 0.2, and the MCQ is considered poorly constructed. If there are two NFDs, the DE is 0.4, and it is regarded as moderate. If there is one NFD, and the MCQ has a DE of 0.8, it is considered as good. If there are no NFDs, and the DE is 1, its construction is regarded as excellent.

Kuder-Richardson Formula 20 (KR-20)

Reliability refers to the consistency of scores across different assessments measuring the same construct (Chen et al., 2020). Assessing the overall reliability of an MCQ assessment is crucial to ensure it consistently produces accurate results (Kumar et al., 2021). Various reliability coefficients, such as Cronbach's alpha, Guttman Split-Half Coefficient (L4), and KR-20 estimate the internal consistency of the MCQ assessment (Triono et al., 2020). Higher coefficients indicate greater reliability and internal consistency. Evaluating assessment reliability allows educators and researchers to gauge the stability of the MCQ results, leading to more confident interpretations and informed decision-making. The KR-20 is calculated by substituting the number of items (k), the fraction of respondents who answered the item correctly (p), the fraction of respondents who answered the item incorrectly (q), and the variance of the total assessment score (σ^2) into Equation 4.

$$KR - 20 = \frac{k}{k-1}\left(1 - \frac{\Sigma pq}{\sigma^2}\right) \tag{4}$$

KR-20 ranges between 0 and 1, with higher values indicating greater reliability. Values above 0.7 are considered highly reliable and suitable for making inferences about individual performance. Values between 0.3 and 0.7 indicate moderate reliability, suggesting some consistency in measuring the intended construct, but revisions or additions to the items may be needed. Values below 0.3 denote low reliability, indicating the assessment does not consistently measure the same construct and may not be reliable for making inferences (Selvi & Özge, 2023).

MCQ Conformity with Design Guidelines (CDG)

Attention to content, format, style, stem, and distractor design ensures valid, reliable, and fair MCQ tests (Haladyna et al., 2019). Content should be specific, avoid overly general material, ensure item independence, and elicit higher-level thinking without trick items. Format items vertically, edit thoroughly, use appropriate language complexity, and minimize unnecessary reading. Stems should be clear, concise, positively worded, and avoid negatives. Distractors should be plausible, free of clues, logically ordered, and avoid humor. Following these guidelines helps create effective MCQ assessments that accurately measure students' knowledge and skills. Higher conformity with the MCQ design guidelines is associated with higher average assessment scores

Results

The present study analyzed 145 MCQ items and 910 responses in three first-year physics modules. The MCQ items comprised a stem, a key and a variable number of distractors. Table 1 summarizes the six assessments analyzed and reports on the number of MCQ items in the assessment, the number of respondents that participated in the assessment, and the percentage of respondents that achieved a score of more than 50%. Assessment XA comprised 20, five-option items; four, four-option items; and one, three-option item. Three hundred and seventy-nine students responded to the assessment.

Table 1. First-year Physics Summative Assessments

Assessment	Items	Students	>50 (%)[1]	Average (%)[2]
XA	25	379	49	49.71
XB	25	165	36	44.36
YA	28	238	38	56.94
YB	27	55	44	48.11
ZA	20	53	83	63.96
ZB	20	20	75	60.00
Total	145	910		

[1]Percentage of respondents that achieved a score of more than 50% correct answers in the assessment. [2]average score of the corresponding assessment.

The analysis of the assessment results focused on the calculated mean score of 12.4, a maximum score of 23 (N = 3), a minimum score of 0 (N = 3), a modal score of 10 (N = 37), and interquartile scores of 9 and 17. The interquartile scores represent the number of students in the high-achieving group (N = 95) and the number of students in the low-achieving group (N = 110). There were 205 students in the high and low achievers' groups combined (N = 205, 95 + 110). The MCQ analysis focused on the calculated DF, DI, and DE for each MCQ item. Question 1 (Q1) of the assessment is used as an example to illustrate the analysis procedure. There are 88 (H) correct responses in the high-achievers group and 32 (L) correct responses in the low-achievers group. Substituting the H (88), L (32), and N (205) values for Q1 into Equations 1 and 2 calculates a DF of 0.32 and a DI of 0.30. Q1 has five options, comprising one key and four distractors (D = 4). None of the Q1 options resulted in a response rate of less than 5%, implying that all of the Q1 distractors were functional (NFD = 0). Substituting D (4) and NFD (0) into Equation 3 calculates a DE of 1. The procedure was repeated for Q2 through to Q25 to calculate the DF, DI, and DE for the respective questions. The analysis of assessment XA is presented in Table 2.

Table 2. Assessment XA MCQ Analysis

Question	Options	H[1]	L[2]	N[3]	DF[4]	DI[5]	NFD[6]	DE[7]
Q1	5	88	32	205	0.59	0.55	0	1.00
Q2	5	82	13	205	0.46	0.67	0	1.00
Q3	5	15	9	205	0.12	0.06	1	0.80
Q4	5	82	18	205	0.49	0.62	1	0.80
Q5	5	76	25	205	0.49	0.50	0	1.00
Q6	5	87	23	205	0.54	0.62	0	1.00
Q7	5	78	16	205	0.46	0.60	0	1.00
Q8	5	91	39	205	0.63	0.51	2	0.60
Q9	5	80	31	205	0.54	0.48	1	0.80
Q10	5	88	23	205	0.54	0.63	0	1.00
Q11	5	88	36	205	0.60	0.51	1	0.80
Q12	5	90	28	205	0.58	0.60	2	0.60
Q13	5	17	37	205	0.26	-0.20	0	1.00
Q14	5	10	26	205	0.18	-0.16	0	1.00
Q15	5	82	29	205	0.54	0.52	1	0.80
Q16	4	87	47	205	0.65	0.39	0	1.00
Q17	5	75	18	205	0.45	0.56	0	1.00
Q18	5	68	24	205	0.45	0.43	0	1.00
Q19	5	86	23	205	0.53	0.61	0	1.00
Q20	5	87	39	205	0.61	0.47	2	0.60
Q21	3	54	27	205	0.40	0.26	0	1.00
Q22	4	85	29	205	0.56	0.55	0	1.00
Q23	4	79	37	205	0.57	0.41	0	1.00
Q24	4	88	37	205	0.61	0.50	0	1.00
Q25	5	68	16	205	0.41	0.51	1	0.80

[1]number of correct answers in high achieving group, [2]number of correct answers in low achieving group, [3]number of respondents in high and low achievers' groups, [4]difficulty index, [5]discrimination index, [6]non-functional distractors, [7]distractor effectiveness.

The data in Table 2 show three inconsistent DF values (Q3, Q13, and Q14) and four inconsistent DI values (Q3, Q13, Q14, and Q21). The data also show 12 NFDs in nine questions (Q3, Q4, Q8, Q9, Q11, Q12, Q15, Q20, and Q25). The reliability calculation required the analysis of each question individually and the overall assessment. The product of the fraction of correct responses (0.58) and incorrect responses (0.42) for Q1 is 0.24. The product of the fraction of correct responses and incorrect responses is repeated for Q2 to Q25. The sum of the calculated products for Q1 to Q25 is divided by the variance (27) of all responses and is used in conjunction with the number of questions (25) to calculate the reliability of the assessment according to Equation 4. The KR-20 for assessment XA is 0.82, indicating a highly reliable assessment. The MCQ analysis procedure to calculate DF, DI, DE, and KR-20 was repeated for the remaining five assessments (Table 3).

Table 3. Mean Difficulty Indices

Assessment	Mean DF[1]	Interpretation
XA	0.49	Ideal difficulty
XB	0.45	Acceptable difficulty
YA	0.58	Ideal difficulty
YB	0.48	Acceptable difficult
ZA	0.54	Ideal difficulty
ZB	0.57	Ideal difficulty

[1]DF = difficulty index. Across all modules $M = 0.52$. $SD = 0.04$.

Discussion

How do the MCQ items' difficulty, discrimination between high and low achievers, suitability of distractors, and overall reliability contribute to the effectiveness of MCQ-based summative assessments?

The data in Table 3 show that the assessments over the two years were acceptably to ideally difficult. The mean DF of all the MCQ items analyzed (M = 0.52, SD = 0.04) represents ideal difficulty. The DF findings of the study are within the range calculated by Sharma (2021), Uddin et al. (2022), and Iqbal et al. (2023). Although the results are in line with those in the literature, there is uncertainty in calculating the DF. Sharma presented two alternate formulas to calculate DF. In the first formula, the total number of correct responses from the higher and lower achieving groups is divided by the sum of respondents in both groups. In the second formula, DF is calculated as the sum of correct responses from the higher and lower-achieving groups, divided by the total number of respondents in the entire assessment.

Iqbal et al. (2023) and Qamar et al. (2022) also used the sum of the number of higher and lower achievers as the denominator of the index. However, Uddin et al. (2022) used the total number of respondents as the denominator of the index. Using a higher number of respondents decreases the DF and it is, therefore, recommended to use the actual number of respondents in each achievement group and not the total number of respondents. Another source of discrepancy may be due to Sharma (2021) using the top 27% and bottom 27% as high and low-achieving groups. Iqbal et al. used the top 33% as the high-achieving groups and the bottom 33% as the low-achieving group. Reza et al. (2021), Uddin et al., and Qamar et al. do not specify how high and low-achieving groups are calculated. This differs from the present study that uses the first and third quartiles to ascertain the high and low achieving groups. The use of quartiles in the present study presents a statistical representation of the data and is, therefore, an acceptable method. The data in Table 4 show that the assessments over the two years ideally discriminated between high achievers and low achievers.

Table 4. Mean Discrimination Indices

Assessment	Mean DI[1]	Interpretation
XA	0.45	Ideal discrimination
XB	0.37	Ideal discrimination
YA	0.50	Ideal discrimination
YB	0.40	Ideal discrimination
ZA	0.41	Ideal discrimination
ZB	0.57	Ideal discrimination

[1]discrimination index. Across all modules M = 0.45. SD = 0.07

The mean DI of all the MCQ items analyzed (M = 0.45, SD = 0.07) represents ideal discrimination. Therefore, the present study finds that the MCQ items of the summative assessments for first-year Physics over the two years appropriately discriminate between high and low achievers. The mean DI of the assessments calculated in the present study are lower than the findings of Sharma (2021), Iqbal et al. (2023), and Uddin et al. (2022), that calculated DFs of 0.58, 0.36, and 0.37 respectively. Similar to the DF calculations, Sharma used the number of respondents in the groups as the denominator of the DI. Iqbal et al. and Qamar et al. (2022) used the same formula as Sharma, while Reza et al. (2021) calculated DI as the difference between correct answers in the high and low-achieving groups. Uddin et al. used the point-biserial method to calculate DI. The present study used the first and third interquartile ranges to calculate the number of respondents in the high and low-achieving groups. The data in Table 5 show that the mean DE of the summative assessments over the two years ranges from moderate to good.

Table 5. Mean Distractor Effectiveness

Assessment	Mean DE[1]	Interpretation
XA	0.90	Good
XB	0.94	Good
YA	0.78	Moderate
YB	0.85	Good
ZA	0.71	Moderate
ZB	0.80	Good

[1]DE = distractor effectiveness. Across all modules M = 0.85. SD = 0.09

The mean DE of all the MCQ items analyzed (M = 0.85, SD = 0.09) represents good distractor effectiveness. Therefore, the present study finds that the MCQ items of the summative assessments for first-year Physics over the two years have suitable distractors. The mean DE calculated in the present study is in alignment with the findings of Sharma (2021) and Reza et al. (2021), who calculated DEs of 0.75 and 0.84, respectively. Iqbal et al. (2023) calculated 100% DE in 34 assessment items. The data in Table 6 show that the reliability of the assessments over the two years ranges from moderately to highly reliable.

Table 6. KR-20 reliability

Assessment	KR-20[1]	Interpretation
XA	0.82	Highly reliable
XB	0.76	Moderately reliable
YA	0.64	Moderately reliable
YB	0.81	Highly reliable
ZA	0.79	Moderately reliable
ZB	0.82	Highly reliable

[1]KR-20 = Kuder-Richardson reliability index. M = 0.81. SD = 0.05.

The mean KR-20 of the MCQ assessments analyzed (M = 0.81, SD = 0.05) represents high reliability. Therefore, the present study finds that the MCQ-based summative assessments for first-year Physics over the two years are highly reliable. The reliability of the assessments calculated in the present study is lower than the findings of Sharma (2021), who calculated assessment reliabilities of 0.97. The findings of the present study are in alignment with Uddin et al. (2022), who calculated reliabilities ranging from 0.69 to 0.86. Sharma used the Guttman Split-Half Coefficient, which is primarily used for two options, while the assessment items were four-

option MCQs. Uddin et al. and Reza et al. (2021) used Cronbach's alpha reliability co-efficient, while Iqbal et al. Iqbal et al. (2023) used the KR-20 reliability co-efficient as used in the present study. The findings of the present study are aligned with the results of Iqbal et al., that calculated KR-20 of 0.75, indicating moderate reliability.

How do the calculated MCQ indices adequately reflect student performance in relation to their conformity with design guidelines and assessment scores?

The data in Table 1 show only 36% of the respondents achieved more than 50%, indicating that 64% of the students did not pass the assessment. The calculated mean DF, DI, DE, and KR-20 represent optimal assessment items. However, the calculated MCQ indices do not adequately reflect the student performance and a qualitative analysis of the individual assessment items presenting unsuitable MCQ statistics is required. The qualitative analysis provides insight into the observed discrepancies between MCQ statistics and student performance and elucidates the factors influencing student outcomes beyond the statistical measures provided. The following Assessment XB items are used as an example to illustrate the general issues uncovered in the qualitative analysis. Questions Q4, Q5, Q14, Q18, and Q23 of Assessment XB presented poor DF and DI.

"Q4: A small planet having a radius of 1000 km exerts a gravitational force of 100 N on an object that is 500 km above its surface. If this object is moved 500 km further from the planet, the gravitational force on it will be closest to:
A. 25 N
B. 71 N
C. 50 N
D. 56 N
E. 75 N"

The conceptual framework provides guidelines for effective MCQ design (Haladyna et al., 2019). According to the guidelines, the stem is well designed as it conveys the central idea clearly and concisely. The distractors form the basis for the poorly calculated DF (0.45) and DI (0.37). The distractors present five options rather than the proposed three options. The distractors are also not presented in a numerical order as the guidelines suggest. The guidelines propose an accurate key, and while the correct answer to the question is 56.25N, the key is rounded down to 56N, without informing the student of the rounding down. The issue of presentation of distractors and rounding of options has been evident in all assessments analyzed. The misalignment of the distractors with the guidelines implies that the MCQ is testing multiple cognitive levels at the same time. The presentation of distractors in an order other than numerical, and the rounding off of options place an additional cognitive burden on the student, which misaligns the MCQ even further. The Spearman rank correlation between the average assessment scores and CDG ($\rho(4) = 0.841$, $p = 0.036$) indicates a statistically significant strong positive correlation suggesting that higher conformity with MCQ design guidelines is associated with higher average assessment scores. Table 7 lists the CDG over the two assessments for each module.

Table 7. MCQ Conformity to Design Guidelines by Module

Assessment	CDG (%)[1]	Score (%)[2]
XA	71	50
XB	64	44
YA	65	57
YB	58	48
ZA	85	57
ZB	87	60

[1]CDG is the average conformity (%) across all MCQ items for each module. [2]average score of all students who completed the assessment for each module.

What are the main characteristics of poorly designed MCQs?

Eleven of the 29 MCQ design guidelines proposed by Haladyna et al. (2019) were infringed upon in more than 10% of the MCQ items analyzed. Table 8 presents the CDG across all MCQ items analyzed.

Table 8. Conformity to Design Guidelines across all MCQs

Design Guideline	CDG (%)[1]
Three options are sufficient	34.00
Place options in logical or numerical order	34.91
Avoid options like none of the above	65.14
Ensure equal decimal places in each option	72.30
Word the options positively	78.33
Vary the location of the right answer	78.98
Base each item on one cognitive level	84.96
Avoid trick items	88.91
Keep the length of options about equal	89.12
State the central idea clearly in the stem	89.39
Avoid pairs or triplets of options	89.86

[1]Conformity with MCQ design guidelines. Guidelines adapted from "Are Multiple-choice Items too Fat?" by T.M. Haladyna, M. C. Rodriguez, and C. Stevens, 2019, Applied Measurement in Education, 32(4), 350–364

The analysis of the CDG reveals key areas of concern in the construction of MCQs. A significant issue is the frequent inclusion of more than three options, which undermines the tests' ability to challenge and differentiate between students effectively. Additionally, the improper ordering of answer choices adds unnecessary confusion, making it harder for students to process information efficiently. These issues highlight a critical need for more rigorous adherence to guidelines that ensure options do not place an additional cognitive burden and are logically presented. Furthermore, the analysis indicates that the use of vague options such as "none of the above" and inconsistent formatting of numerical options are common pitfalls. Such practices can compromise the reliability of assessments by allowing students to guess answers rather than demonstrating true understanding. The presence of negative wording and predictable answer patterns further complicates the assessment process, potentially misleading students and diminishing the overall fairness of the test (Alonzo et al., 2023). These infringements suggest that test designers often overlook the importance of clarity and consistency in question-and-answer construction.

In addition to these major concerns, the analysis also points to subtler issues like the inclusion of trick items and the failure to base each question on a single type of content. These factors can confuse and frustrate students, leading to results that do not accurately reflect their knowledge or skills. Ensuring that options are of equal length, central ideas are clearly stated, and avoiding clueing pairs or triplets are all crucial for maintaining the integrity of the assessment. Addressing these infringements is essential for developing MCQs that are both fair and effective in measuring student performance.

Conclusion

Three first-year university Physics modules were analyzed over the two years, and the MCQs used as an assessment tool were analyzed. The analysis involved calculating and interpreting the DF, DI, DE, KR-20, and CDG for the MCQs. The findings of the present study indicate that while the MCQ assessments marginally discriminated between high and low achievers, and some questions were too difficult. Problems creating the observed difficulty include low distractor effectiveness and poor design. The presence of NFDs caused respondents to guess the correct answer, which did not fulfil the aims of the assessment. This observation is highlighted by the low calculated assessment reliability. Higher conformity with the MCQ design guidelines is associated with higher average assessment scores.

References

Adnan, S., Sarfaraz, S., Nisar, M. K., & Jouhar, R. (2023). Faculty perceptions on one-best MCQ development. *The Clinical Teacher*, *20*(1). https://doi.org/10.1111/tct.13529

Ahmad, S. (2019). Analysis of test items used in an achievement test in physics at secondary level. *Journal of Education and Practice*, *10*(10), 90–96. https://doi.org/https://doi.org/10.7176/JEP

Al-zboon, H. S. (2022). The impact of the number of distractors in multiple-choice test items on the psychometric characteristics of the items and item information function according to the two-parameter logistic model in the item response theory. *Journal of Education and Practice*, *13*(13), 53–64. https://doi.org/10.7176/JEP/13-13-07

Alkhatib, O. J. (2022). An effective assessment method of higher-order thinking skills in engineering and humanities. *2022 Advances in Science and Engineering Technology International Conferences (ASET)*, 1–6. https://doi.org/10.1109/ASET53988.2022.9734856

Alonzo, D., Bejano, J., & Labad, V. (2023). Alignment between teachers' assessment practices and principles of outcomes-based education in the context of Philippine education reform. *International Journal of Instruction*, *16*(1), 489–506. https://e-iji.net/ats/index.php/pub/article/view/205

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*(4), 581–594. https://doi.org/10.1177/014662167800200413

Ang, K. C., & Boo, H. K. (2006). Exploring the use of reflective MCQ in primary science formative assessment. In Y. J. Lee, A. L. Tan, & B. T. Ho (Eds.), *Science education: What works* (pp. 86–97). National Institute of Education. https://repository.nie.edu.sg/handle/10497/14904

Ansari, M., Sadaf, R., Akbar, A., Rehman, S., Chaudhry, Z. R., & Shakir, S. (2022). Assessment of distractor efficiency of MCQS in item analysis. *The Professional Medical Journal*, *29*(05), 730–734. https://doi.org/10.29309/TPMJ/2022.29.05.6955

Anunpattana, P., Khalid, M. N. A., & Iida, H. (2023). Objectivity and subjectivity in variation of multiple choice questions: Linking the theoretical concepts using motion in mind. *IEEE Access*, *11*, 35371–35397. https://doi.org/10.1109/ACCESS.2023.3265196

Baldwin, P., Mee, J., Yaneva, V., Paniagua, M., D'Angelo, J., Swygert, K., & Clauser, B. E. (2022). A natural-language-processing-based procedure for generating distractors for multiple-choice questions. *Evaluation & the Health Professions*, *45*(4), 327–340. https://doi.org/10.1177/01632787211046981

Chen, Q., Zhu, G., Liu, Q., Han, J., Fu, Z., & Bao, L. (2020). Development of a multiple-choice problem-solving categorization test for assessment of student knowledge structure. *Physical Review Physics Education Research*, *16*(2), 1–10. https://doi.org/10.1103/PhysRevPhysEducRes.16.020120

Chit, Y. Z. (2020). An analysis on functioning and non functioning distractors in physics multiple choice question. *International Asian Congress on Contemporary Sciences - IV*, 218–227. https://www.researchgate.net/publication/344453592

Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, *12*(1), 19–24. https://doi.org/10.1016/S1322-7696(08)60478-3

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement*, *16*(1), 13–37. https://doi.org/10.1177/001316445601600102

Dwivedi, Y. K., Hughes, D. L., Coombs, C., Constantiou, I., Duan, Y., Edwards, J. S., Gupta, B., Lal, B., Misra, S., Prashant, P., Raman, R., Rana, N. P., Sharma, S. K., & Upadhyay, N. (2020). Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life. *International Journal of Information Management*, *55*(December 2020), 102211. https://doi.org/10.1016/j.ijinfomgt.2020.102211

Fadzil, H. M., Saat, R. M., & Rafi, A. (2022). Development of technology-enhanced three-tier diagnostic test to assess pre-university students' understanding of scientific concepts. In *Alternative Assessments in Malaysian Higher Education* (pp. 285–292). Springer Singapore. https://doi.org/10.1007/978-981-16-7228-6_29

Haladyna, T. M., Rodriguez, M. C., & Stevens, C. (2019). Are Multiple-choice Items Too Fat? *Applied Measurement in Education*, *32*(4), 350–364. https://doi.org/10.1080/08957347.2019.1660348

Hassen, H. (2022). How ethiopian standardized national examinations achieve their goal? 2014/15 university entrance examination exam in focus. *African Journal of Social Sciences and Humanities Research*, *5*(3), 1–14. https://doi.org/10.52589/AJSSHR-PBV0DRVO

Ijiwade, O., & Alonzo, D. (2023). Teacher perceptions of the Use of a computer-adaptive test for formative purposes: typologies of practices. *International Journal of Instruction*, *16*(2), 887–908. https://doi.org/10.29333/iji.2023.16247a

Iqbal, Z., Saleem, K., & Arshad, H. M. (2023). Measuring teachers' knowledge of student assessment: Development and validation of an MCQ test. *Educational Studies*, *49*(1), 166–183.

https://doi.org/10.1080/03055698.2020.1835615

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. https://doi.org/https://doi.org/10.33225/pec/20.78.576

Jiang, Z., Ouyang, J., Li, L., Han, Y., Xu, L., Liu, R., & Sun, J. (2022). Cost-effectiveness analysis in performance assessments: A case study of the objective structured clinical examination. *Medical Education Online*, *27*(1), 1–7. https://doi.org/10.1080/10872981.2022.2136559

Kucwaj, H., Ociepka, M., & Chuderski, A. (2022). Various sources of distraction during analogical reasoning. *Memory & Cognition*, *50*(7), 1614–1628. https://doi.org/10.3758/s13421-022-01285-3

Kumar, A. P., Nayak, A., Shenoy, M. K., Goyal, S., & Chaitanya, M. (2023). A novel approach to generate distractors for Multiple Choice Questions. *Expert Systems with Applications*, *225*, 120022. https://doi.org/10.1016/j.eswa.2023.120022

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical Journal Armed Forces India*, *77*, S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.007

Lawley, D. N. (1943). XXIII.—On Problems connected with Item Selection and Test Construction. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences*, *61*(3), 273–287. https://doi.org/10.1017/S0080454100006282

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). Princeton University Press.

Lord, F. M. (1956). The measurement of Growth. *ETS Research Bulletin Series*, *1956*(1). https://doi.org/10.1002/j.2333-8504.1956.tb00058.x

Mendez-Carbajo, D. (2023). The effectiveness of logical distractors in an online module. *Eastern Economic Journal*, *49*(1), 15–30. https://doi.org/10.1057/s41302-022-00232-z

Mittelhaëuser, M., Béguin, A. A., & Sijtsma, K. (2015). The Effect of Differential Motivation on IRT Linking. *Journal of Educational Measurement*, *52*(3), 339–358. https://doi.org/10.1111/jedm.12080

Nojomi, M., & Mahmoudi, M. (2022). Assessment of multiple-choice questions by item analysis for medical students' examinations. *Research and Development in Medical Education*, *11*(24), 1–6. https://doi.org/10.34172/rdme.2022.024

Peeters, M., Cor, M. K., & Maki, E. (2021). Providing validation evidence for a Clinical-science module: Improving testing reliability with quizzes. *Innovations in Pharmacy*, *12*(1), 1–5. https://doi.org/10.24926/iip.v12i1.2235

Popham, W. J. (1990). *Modern educational measurement : a practitioner's perspective* (2nd ed.). Prentice-HallEnglewood Cliffs.

Qamar, A. M., Kanwal, W., & Nadeem, H. A. (2022). Item analysis for test to examine the effect of e-module on the academic perfromance of 7th class science students in Islamabad. *Jahan Tahqee*, *5*(2), 188–196.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests [monograph]*. Nielsen & Lydiche.

Reza, M., Puspita, K., & Oktaviani, C. (2021). Quantitative analysis towards higher order thinking skills of chemistry multiple choice questions for university admission. *Jurnal IPA & Pembelajaran IPA*, *5*(2), 172–185. https://doi.org/10.24815/jipi.v5i2.20508

Selvi, H., & Özge, A. (2023). Reliability and Validity of Clinical Scales Measurement. In P. Y. Dikmen & A. Özge (Eds.), *Clinical Scales for Headache Disorders* (pp. 45–60). Springer. https://doi.org/10.1007/978-3-031-25938-8_3

Sharma, L. R. (2021). Analysis of difficulty index, discrimination index and distractor efficiency of multiple choice questions of speech sounds of english. *International Research Journal of MMC*, *2*(1), 15–28. https://doi.org/10.3126/irjmmc.v2i1.35126

Siddiqui, S. (2022). Categorized and correlated multiple-choice questions: A tool for assessing comprehensive physics knowledge of students. *Education Sciences*, *12*, 1–19. https://doi.org/10.3390/educsci12090575

Sideris, G. A., Singh, A., & Catanzano, T. (2022). Writing high-quality multiple-choice questions. In *Image-Based Teaching* (pp. 123–146). Springer International Publishing. https://doi.org/10.1007/978-3-031-11890-6_9

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika*, *52*(1), 79–97. https://doi.org/10.1007/BF02293957

Smith, T. I., Louis, K. J., Ricci, B. J., & Bendjilali, N. (2020). Quantitatively ranking incorrect responses to multiple-choice questions using item response theory. *Physical Review Physics Education Research*, *16*(1), 1–16. https://doi.org/10.1103/PhysRevPhysEducRes.16.010107

Tatsuoka, K. . (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354. http://www.jstor.org/stable/1434951

Tolba, E. G. A., & Youssef, N. H. (2024). Conceptual change and developing mental motivation in physics: Effects of transformational learning theory. *International Journal of Instruction*, *17*(4), 359–384. https://doi.org/10.29333/iji.2024.17421a

Tollefson, N. (1987). A comparison of the item difficulty and item discrimination of multiple-choice items using the "None of the Above" and one correct response Options. *Educational and Psychological Measurement*, *47*(2), 377–383. https://doi.org/10.1177/0013164487472010

Triono, D., Sarno, R., & Sungkono, K. R. (2020). Item analysis for examination test in the postgraduate student's selection with Classical Test Theory and Rasch Measurement Model. *2020 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 523–529. https://doi.org/10.1109/iSemantic50169.2020.9234204

Uddin, M. K., Parvez, R. A., Mullick, T. T., & Habib, M. A. (2022). Multiple choice questions in higher secondary examination in bangladesh: a comparative evaluation by year and stream. *International Journal of Asia Pacific School Psychology*, *3*(1), 71–82. https://www.researchgate.net/publication/358668806

Wright, B. D. (1968). Sample-free test calibration and person measurement. *National Seminar on Adult Education Research*. https://eric.ed.gov/?id=ED017810

Xu, L., Jin, R., Huang, F., Zhou, Y., Li, Z., & Zhang, M. (2020). Development of Computerized Adaptive Testing for Emotion Regulation. *Frontiers in Psychology*, 1–11. https://doi.org/10.3389/fpsyg.2020.561358