Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.17, No.4, 2026

Sensitivities of Multiple-Choice and Multi-Item Mirror Question Tests in Determining Misassessment Risk Among First-Year Medical Students in Foundation Biology and Physics at Levy Mwanawasa Medical University

Ephraim Chongo *, Chipasha Kaminsa, Mirriam Kaona and Mathias Chamatwa Zulu

Institute of Basic and Biomedical Science (IBBS), Natural Sciences Department – Levy Mwanawasa Medical University, P. O Box 33991, Lusaka, Zambia.

*E-mail of the Corresponding Author: ephraim.chongo@lmmu.ac.zm

Abstract

This study examined the extent to which traditional Number-Right Multiple-Choice Question (MCQ) scoring misassesses students' true knowledge compared with the Multi-Item Mirror Question Test (MIMQT) model, which uses a Knowledge Equivalence Scoring (KES) system that awards equal value for identifying what applies and what does not apply within a concept. A comparative quasi-experimental design was employed, involving 119 first-year Foundation Biology and Physics students at Levy Mwanawasa Medical University, randomly assigned to either a traditional MCQ test (n = 58) or an equivalent-content MIMQT model (n = 61). Performance scores and student perceptions were collected, with the latter measured through a reliable questionnaire (α = .876). Results showed significantly higher performance under the MIMQT model in both Physics (Mean = 7.62 vs. 6.38) and Biology (Mean = 11.16 vs. 8.45). An independent samples t-test confirmed a statistically significant difference between scoring models, t(94) = –3.58, p = .001. In this study, traditional MCQs mismeasured 17% of Physics and 24% of Biology knowledge, meaning that MIMQT model improved measurement by the same percentages. Perception data revealed strong preference to the MIMQT model across fairness, accuracy, motivation, and reduced misassessment, with 75%–95% of respondents agreeing on key components. Overall, the findings indicate that the MIMQT model provides a more accurate, equitable, and diagnostically sensitive measure of true knowledge than traditional MCQs by assessing bidirectional understanding with KES rather than unidirectional single-answer recognition with "all" or "nothing" scoring system. The findings show that traditional Number-Right MCQs do not fully capture true knowledge because they rely on selecting a single best answer and ignore the knowledge of correctly identifying incorrect options. Since these formats operate as disguised True/False systems without awarding credit for knowing what does not apply, they provide a narrow and unidirectional judgment of understanding. In contrast, the MIMQT model, which scores both forms of identifications with equal value, offers a more comprehensive measurement of true knowledge (including peripheral knowledge) by recognising bidirectional understanding of a concept. If this was not the case, only "true" responses would earn credit while correct recognition of "false" would not earn points and dismissed as non-knowledge.

Keywords: MIMQT Model, Knowledge Equivalence Scoring, Multiple-Choice Questions, Misassessment Risk, True Knowledge Measurement, Partial Knowledge, MCQ sensitivity, Number – Right Scoring

DOI: 10.7176/JEP/17-4-10
Publication date: April 30th 2026

1.0: Background to the Study

Recent literature has highlighted that although MCQs are widely favored in healthcare education for their practicality and standardization, their effectiveness in promoting clinical competence and fairness is conditional (Parekh & Bahadoor, 2024) as benefits largely depend on the quality of question design, the desired competence level, and the control of confounding factors such as differential attainment. In the teaching-learning environment, there is a constant need to gauge the quantity or quality of responsiveness of the teaching and learning process, a crucial symbiotic process generally referred to as assessment. Measurement is undertaken to quantify the level of knowledge or skills acquired by a learner (Adom et al., 2020). At the end of every examination (assessment), the examinees' responses have to be analyzed and scored to derive information about examinees' true knowledge (Kanzow et al., 2023), that is accurately quantifying the knowledge. This study on

multiple-choice assessment methods identified 21 different scoring approaches from 258 sources, revealing that examination results varied significantly depending on the method used. In exams with single-choice items, the scores did not always reflect the examinees' true knowledge, highlighting the need to consider both the scoring method and the number of answer options per item when interpreting scores and setting pass marks (Kanzow et al., 2023). The study further showed that the way exam questions are scored can greatly affect the results. For example, if a student truly knows 50% of the material, their final score can still vary a lot depending on the scoring method and the number of answer choices per question. For questions with 2 answer options, the student's score could be as low as 0% or as high as 87.5%, even though they only knew half the content. For questions with 5 answer options, the score could be as low as –60% (a penalty for wrong answers) or as high as 92%. This means that two students with the same level of knowledge could get very different scores, just because of how the test is scored and how many choices each question has. We can point out also here that two students both scoring equal marks (or zero points) on the same question have different levels of understanding. These findings show that MCQs have potential to give an inaccurate reflection of what the student knows and hence the results can be false positive, false negative or can give misleading equivalence. This kind of assessment is more of a professional task than academic exercise for measuring learning. The all-or-nothing scoring of MCQ-based assessments has a major shortcoming because it fails to capture the partial understanding that an examinee may have (Schneid et al., 2025) and it is difficult to understand what knowledge is worth zero score. Such a testing method has risk of Misassessment, which refers to the probability that the assessment gives an inaccurate reflection of what the student knows. The questions to ask are: is it the test tool (MCQ) failing to get the true knowledge from a testee or the testee knows nothing? How far is the zero mark from the true knowledge one has on a particular question? How much is the risk of mis-assessing using MCQ? Are current MCQs scoring systems accurate? In a study involving 105 Type A multiple-choice questions (MCQs), partial credit was awarded on 31 questions, representing 30% of the total. When partial credit was applied, the average score on these questions increased from 76.5% to 86.1%, a clear improvement of 9.6 percentage points was recorded. This difference was statistically significant ($P < .0001$) based on a paired Student t-test (Schneid et al., 2025). This means that examinees are robbed of points when existing MCQs scoring systems are used and the performance reported is misleading. Educators fail to accurately measure learning of their students and learners together with all stakeholders are misinformed about academic performance.

A test, including MCQ test format is used to assess and measure mastery of a skill or knowledge in a particular course of field. And so, it must be valid and reliable in accurately measuring knowledge on questions of interest. While Number Right MCQ tests have several advantages, it has shortcomings as well, particularly in scoring the knowledge, where it is vague and meaningless as to what knowledge zero mark means.

This study aims at determining the magnitude of Misassessment caused by traditional Number Right MCQ scoring system when it is compared to Multi-Item Mirror Question Test (MIMQT) Model's Bidirectional Knowledge Scoring (BKS) on the academic performance in First-Year Foundation Biology and Physics at Levy Mwanawasa Medical University**.**

1.1: Statement of the Problem

Despite the ubiquitous use of multiple-choice questions (MCQs) in assessing knowledge across medical education (Steele et al., 2025), current scoring systems, particularly the number-right and negative marking methods, present serious concerns regarding their validity and fairness as examinees and examiners alike frequently perceive MCQ-based testing as 'unfair (McCoubrie, 2004) and has been asking for improvements from inception. These methods treat guessing as malpractice, yet offer no mechanism to distinguish informed reasoning from random selection and has never been mentioned as a form of examination malpractice in institutional examination malpractice documents and in studies on the vice. Worse still, number-right and negative marking methods discretize knowledge into binary outcomes, "right or wrong", (meaning 1 or 0), is an oversimplification without accommodating the spectrum of partial knowledge that many students possess (Desy et al., 2024). A correct answer scores one point, while a wrong answer or no attempt scores zero, making it vague and academically unjustifiable what type of knowledge merits zero. This all-or-nothing approach leads to misassessment, where students with partial knowledge are equated with those who know nothing, distorting the evaluation of learning, where

1. false positive: partial knowledge scored as full mark due to discretization and dichotomizing of knowledge "all – or nothing", an NR MCQs inherent principle.
2. false negative: student knows partially or fully but scores 0.
3. misleading equivalence: two students both score 1 or 0 mark but have different levels of understanding

(Desy et al., 2024).

On the other hand, partial credit scoring system models attempt to either ordinalize or nominalize knowledge by allocating scores based on proximity to the correct answer. However, these too face limitations, particularly the subjectivity in assigning partial marks and the lack of interrater reliability in determining which option is "closer" to the truth and by how much. This lack of objectivity undermines the reliability of assessment outcomes and risks masking true academic performance with artificial scores. The inability of current common MCQ models to accurately and fairly measure the gradient of knowledge contributes to a significant risk of misassessment, where reported scores fail to reflect actual quantification of true knowledge in a learner. As such, there is an urgent need to evaluate the magnitude of misassessment and identify a scoring approach that can more reliably quantify learners' true knowledge. This study proposes a Multi-Item Mirror Question Test model with Knowledge Equivalence Scoring (KES) system to extract true knowledge from the learner. This scoring system can also be called Bidirectional Knowledge Scoring (BKS) or Symmetric Knowledge Scoring (SKS).

1.2: Research Questions

1. To what extent does the KES MIMQT model yield different mean performance outcome compared to NR MCQ scoring model?
2. How do learners perceive the MIMQT model versus conventional scoring method?
3. What scoring method produces lower risk of Misassessment between Number Right MCQ and Knowledge Equivalence MIMQT scoring systems on academic performance in First-Year Foundation Biology and Physics Courses at Levy Mwanawasa Medical University?

1.3: Hypotheses

Null Hypothesis H_o: $\bar{x}\ Traditional\ MCQ\ Mean\ Score = \bar{x}\ MIMTQ\ Mean\ Score$

Alternative Hypothesis H_a: $\bar{x}\ Traditional\ MCQ\ Mean\ Score \neq \bar{x}\ MIMTQ\ Mean\ Score$

2.0: METHODOLOGY

2.1: Research Design

This study employed a comparative quasi-experimental research design intended to evaluate the effectiveness, fairness, and accuracy of two assessment models: the traditional Number-Right (NR) Multiple-Choice Question (MCQ) scoring system and the Knowledge Equivalence Scoring (KES) of the Multi-Item Mirror Question Test (MIMQT) model. Two independent groups of students were assessed with parallel tests covering identical content but scored with different approaches. This design allowed for a direct comparison of performance outcomes, student perceptions, and levels of misassessment associated with each assessment model.

2.2: Population and Sample

The study population comprised First-Year Foundation students enrolled in Foundation Biology and Foundation Physics at Levy Mwanawasa Medical University (N= 645). The sample consisted of 119 students, with random assignments of 58 to the control group, which used traditional MCQs, and 61 to the experimental group, which used the MIMQT model.

2.3: Instruments

1. **Test**

Two parallel equivalent test models were administered to the two groups. The control group completed a 30-mark traditional MCQ test consisting of 15 items from Biology and 15 from Physics. Each question awarded one mark for a correct response and zero for an incorrect or no response. The experimental group completed a 30-mark test constructed using the MIMQT model. Like the MCQ test, it consisted of 15 Biology items and 15 Physics items, but each question contained four statements to be ticked if correct and crossed if incorrect. Each option carried 0.25 marks, allowing each question to total one mark. Every MIMQT item was carefully designed to mirror the content and intent of the traditional MCQ counterpart to ensure equivalence in content coverage and level of cognitive engagement.

Example from the Control Group:

Which of the following is true about "**a**" and "**b**" on vectors and scalars?

A. They are two scalars

B. They are two vectors

C. They are two coordinates of a scalar
D. They are two angles

Example from the Experimental Group:

The following are vector notations. For each option **Tick** if correct and **Cross** if incorrect.

A. **b** [ ]
B. $\overline{AB}$ [ ]
C. $\rightarrow$ [ ]
D. $\vec{p}$ [ ]

Such a question assesses understanding of vector notation more comprehensively and accurately than the one in the control group. Every question in the MIMQT model was a mirror question from the control group with traditional MCQs, hence they were equivalent tests.

2. Questionnaire

A structured questionnaire was administered to gather students' perceptions of the assessment model they used. It measured perceptions related to fairness in scoring, recognition of partial knowledge, motivation to attempt all items, transparency of the model, accuracy in reflecting true knowledge, and reduction of guessing and misassessment. Responses were captured using a five-point Likert scale. The instrument demonstrated high reliability, with a Cronbach alpha of 0.876.

2.4: Procedure

The procedure began with the administration of the assessment instruments to the two groups. Students in the control group wrote the traditional MCQ test, while those in the experimental group wrote the MIMQT test. After completing the tests, the scripts were scored according to the respective scoring models, with MCQ scripts scored dichotomously and MIMQT scripts scored using Knowledge Equivalence Scoring (KES) system which also can be termed as Bidirectional Knowledge Scoring (BKS) or Symmetric Knowledge Scoring SKS). The perception questionnaire was at the end of the experimental group test paper. It was answered by participants in the experimental group only. The test scores and questionnaire responses were then compiled and coded for analysis.

2.5: Data Collection

Data were collected from two primary sources: the performance scores obtained from the MCQ and MIMQT assessments, and the responses to the perception questionnaire. The test scores provided quantitative evidence on the performance differences between the two groups, while the questionnaire responses provided qualitative insights into how students perceived the fairness, accuracy, and overall experience of the assessment models.

3.0: Results and analysis

The results for research question 1 for both foundation Physics and Biology are presented below.

Table 1: Foundation Physics descriptive statistics

Physics Groups		N	Mean	Std. Deviation	Std. Error Mean
Physics Score	MCQs	58	6.38	2.26	.30
	MIMQT	61	7.62	1.39	.18

The descriptive analysis revealed that the MIMQT model achieved a higher performance, with a mean score of 7.62 (50.8%), a standard deviation of 1.39, and a standard error of 0.18. In comparison, the traditional MCQs model recorded a lower mean score of 6.38 (42.5%), with a standard deviation of 2.26 and a standard error of 0.30 in Foundation Physics.

Table 2: Foundation Biology descriptive statistics

Biology Groups		N	Mean	Std. Deviation	Std. Error Mean
Biology Score	MCQs	58	8.45	2.49	.33
	MIMQT	61	11.16	1.62	.21

For Foundation Biology, the MIMQT model recorded a mean score of 11.16 (74.40%), with a standard deviation of 1.62 and a standard error of the mean (SEM) of 0.21. In comparison, the traditional MCQs model achieved a lower mean score of 8.45 (56.30%), accompanied by a standard deviation of 2.49 and a standard error of the mean of 0.33.

An independent samples t-test was conducted to compare the mean scores of the Traditional MCQ model and the MIMQT model. Levene's test for equality of variances was significant ($p = 0.001$), indicating that the assumption of equal variances was violated; therefore, the unequal-variance was assumed.

The results showed a significant difference between the two mean scores, $t_{(94)} = -3.58$, $p = 0.001$. Thus, the null hypothesis (H_0: the mean score of the Traditional MCQ model equals the mean score of the MIMQT model) was rejected, and the alternative hypothesis (H_a: the two mean scores are not equal) was supported.

With all 61 participants indicating prior experience with traditional NR MCQs, results presented in Tables 3-14 were used to address Research Question 2, which examined students' perceptions of the MIMQT model with regard to scoring fairness, motivation and confidence, accuracy in measuring true knowledge, and the risk of misassessment. The analysis also explored the overall fairness of the model in assessing true knowledge and allocating scores.

Table 3: Gender Vs (Perception on Rewarding partial knowledge more fairly than usual MCQs) crosstab

Fairness on rewarding partial knowledge						Total
Gender	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	
Male	1	0	1	5	13	20
Female	2	4	6	11	18	41
Total	3	4	7	16	31	61

Most males, 13 strongly agreed and 5 agreed that the model rewards partial knowledge more fairly, giving a total of 18 out of 20 males (90%) with a positive perception. Only 1 male was neutral and 1 disagreed (5% each), indicating strong confidence. Among females, 18 strongly agreed and 11 agreed, totaling 29 out of 41 (71%) with a positive perception. Six were neutral and 4 disagreed (15% each), while 2 strongly disagreed (5%), showing slightly more varied opinions than males. Overall, 31 strongly agreed and 16 agreed, totaling 47 out of 61 respondents (77%) with a positive perception. Seven were neutral (11.5%) and 7 disagreed or strongly disagreed (11.5%), indicating that most respondents perceive the model as fairer in rewarding partial knowledge, with males showing slightly higher confidence than females.

Table 4: Gender vs (Chances of being unfairly marked when I know something) crosstab

Gender	Strongly Disagree	Disagree	Agree	Strongly Agree	Total
Male	1	0	4	15	20
Female	3	1	11	26	41
Total	4	1	15	41	61

The majority of male respondents, 15 strongly agreed and 4 agreed that the model reduces the chances of being

unfairly marked, resulting in 19 out of 20 males (95%) expressing a positive perception, with no males being neutral and only 1 (5%) strongly disagreeing. Among females, 26 strongly agreed and 11 agreed, totaling 37 out of 41 (90%) with a positive perception, while 3 strongly disagreed and 1 disagreed (7.3% and 2.4%, respectively), and none were neutral. Overall, 41 respondents strongly agreed and 15 agreed, making 56 out of 61 (91.8%) with a positive perception, with the remaining 5 (8.2%) disagreeing or strongly disagreeing, showing strong confidence in the model's fairness and highlighting that no respondents were undecided on this component.

Table 5: Gender vs (Model makes guessing less likely) crosstab

Gender	Strongly Disagree	Neutral	Agree	Strongly Agree	Total
Male	3	2	7	8	20
Female	1	9	15	15	40
Total	4	11	22	23	60

Among males, 8 strongly agreed and 7 agreed that the component is positively perceived, giving a total of 15 out of 20 males (75%) with a positive perception. Two males were neutral (10%), and 3 strongly disagreed (15%), showing that while most males were positive, a small portion were undecided or disagreed.

Among females, 15 strongly agreed and 15 agreed, totaling 30 out of 40 females (75%) with a positive perception. Nine females were neutral (22.5%), and 1 strongly disagreed (2.5%), indicating that while agreement was high, a larger proportion of females were undecided compared to males.

Overall, 23 respondents strongly agreed and 22 agreed, giving 45 out of 60 respondents (75%) with a positive perception. Eleven respondents were neutral (18.3%), and 4 (6.7%) disagreed, showing that many respondents view this component positively, although some remain undecided.

Table 6: Gender vs (understanding of the scoring system) crosstab

Gender	Disagree	Neutral	Agree	Strongly Agree	Total
Male	0	2	10	8	20
Female	2	7	15	17	41
Total	2	9	25	25	61

Among males, 8 strongly agreed and 10 agreed that the component is positively perceived, totaling 18 out of 20 males (90%) with a positive perception. Two males were neutral (10%), and none disagreed, showing strong overall confidence with no active opposition.

Among females, 17 strongly agreed and 15 agreed, totaling 32 out of 41 females (78%) with a positive perception. Seven females were neutral (17%), and 2 disagreed (5%), indicating that while most females were positive, a notable portion were undecided or slightly opposed.

Overall, 25 respondents strongly agreed and 25 agreed, giving 50 out of 61 respondents (81.9%) with a positive perception. Nine respondents were neutral (14.8%), and 2 (3.3%) disagreed, showing that the majority of respondents view this component positively, though some remain undecided.

Table 7: Gender vs (Model is transparent and easy to follow) crosstab

Gender	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Male	2	2	1	6	9	20
Female	1	8	7	13	12	41
Total	3	10	8	19	21	61

Regarding the perception that the model is transparent and easy to follow, among males, 9 strongly agreed and 6 agreed, totaling 15 out of 20 males (75%) with a positive perception. One male was neutral (5%), while 2 strongly disagreed and 2 disagreed (10% each), showing that most males found the model transparent, though a small portion had reservations.

Among females, 12 strongly agreed and 13 agreed, totaling 25 out of 41 females (61%) with a positive perception. Seven were neutral (17%), 8 disagreed (19.5%), and 1 strongly disagreed (2.5%), indicating that while a majority of females perceived the model as transparent and easy to follow, there was a wider range of opinions compared to males.

Overall, 21 respondents strongly agreed and 19 agreed, giving 40 out of 61 respondents (65.6%) with a positive perception. Eight respondents were neutral (13.1%), and 10 (16.4%) disagreed or strongly disagreed, showing that most respondents perceive the model as transparent and easy to follow, though some expressed uncertainty or disagreement.

Table 8: Gender vs (Motivates me to attempt all items) crosstab

Gender	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Male	1	0	2	8	9	20
Female	1	2	2	16	20	41
Total	2	2	4	24	29	61

Regarding the perception that the model motivates me to attempt all items, among males, 9 strongly agreed and 8 agreed, totaling 17 out of 20 males (85%) with a positive perception. Two males were neutral (10%), and 1 strongly disagreed (5%), indicating that most males felt motivated by the model.

Among females, 20 strongly agreed and 16 agreed, totaling 36 out of 41 females (87.8%) with a positive perception. Two females were neutral (4.9%), and 3 (7.3%) disagreed or strongly disagreed, showing that the majority of females were highly motivated, with very few expressing doubt.

Overall, 29 respondents strongly agreed and 24 agreed, giving 53 out of 61 respondents (86.9%) with a positive perception. Four respondents were neutral (6.6%), and 2 (3.3%) disagreed or strongly disagreed, indicating that the model is largely seen as motivating respondents to attempt all items.

Table 9: Gender vs (Confident that it reflects my true knowledge) crosstab

Gender	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Male	1	0	0	7	11	19
Female	3	3	4	11	19	40
Total	4	3	4	18	30	59

Regarding the perception that the model reflects my true knowledge, among males, 11 strongly agreed and 7 agreed, totaling 18 out of 19 males (94.7%) with a positive perception. No males were neutral, and 1 (5.3%) strongly disagreed, showing very high confidence among male respondents.

Among females, 19 strongly agreed and 11 agreed, totaling 30 out of 40 females (75%) with a positive perception. Four females were neutral (10%), 3 disagreed (7.5%), and 3 strongly disagreed (7.5%), indicating that while most females were confident, some expressed uncertainty or doubt.

Overall, 30 respondents strongly agreed and 18 agreed, giving 48 out of 59 respondents (81.4%) with a positive perception. Four respondents were neutral (6.8%), and 7 (11.8%) disagreed or strongly disagreed, showing that most respondents believe the model accurately reflects their true knowledge, though a minority expressed reservations.

Table 10: Gender vs (Model More accurately measures my true knowledge) crosstab

Gender	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Male	3	0	1	6	10	20
Female	2	3	3	14	19	41
Total	5	3	4	20	29	61

Regarding the perception that the model is a more accurate measurement tool of my true knowledge, among males, 10 strongly agreed and 6 agreed, totaling 16 out of 20 males (80%) with a positive perception. One male was neutral (5%), and 3 strongly disagreed (15%), indicating that most males found the model accurate, though some expressed disagreement.

Among females, 19 strongly agreed and 14 agreed, totaling 33 out of 41 females (80.5%) with a positive perception. Three were neutral (7.3%), 3 disagreed (7.3%), and 2 strongly disagreed (4.9%), showing that while the majority of females perceived the model as accurate, a small portion were neutral or disagreed.

Overall, 29 respondents strongly agreed and 20 agreed, giving 49 out of 61 respondents (80.3%) with a positive perception. Four respondents were neutral (6.6%), and 8 (13.1%) disagreed or strongly disagreed, indicating that most respondents view the model as a more accurate measure of their true knowledge, though some expressed reservations.

Table 11: Gender vs (Model reduces chances of my knowledge being misrepresented) crosstab

Gender	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Male	0	1	1	5	13	20
Female	2	1	9	11	18	41
Total	2	2	10	16	31	61

Most male respondents perceived that the model reduces the chances of their knowledge being misrepresented, with 13 strongly agreeing and 5 agreeing, totaling 18 out of 20 males (90%) showing a positive perception. One male was neutral (5%) and 1 disagreed (5%), indicating strong confidence overall.

For females, 18 strongly agreed and 11 agreed, giving 29 out of 41 females (70.7%) with a positive perception. Nine were neutral (22%), 1 disagreed (2.4%), and 2 strongly disagreed (4.9%), reflecting that while most females were positive, a notable portion were undecided or slightly opposed.

Looking at all respondents, 31 strongly agreed and 16 agreed, totaling 47 out of 61 respondents (77%) with a positive perception. Ten respondents were neutral (16.4%), and 4 (6.6%) disagreed or strongly disagreed, showing that overall, the majority of respondents view the model as reducing the chances of their knowledge being misrepresented, though some expressed uncertainty or disagreement.

Table 12: Gender vs (Model Reduces false positive/false negative/false equivalence) crosstab

Gender	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Male	1	0	4	8	7	20
Female	3	3	10	13	12	41
Total	4	3	14	21	19	61

More male respondents indicated that the model reduces false positives, false negatives, and false equivalence, with 7 strongly agreeing and 8 agreeing, totaling 15 out of 20 males (75%) with a positive perception. Four males were neutral (20%), and 1 (5%) strongly disagreed, showing that while most males were positive, some were undecided.

Among females, 12 strongly agreed and 13 agreed, giving 25 out of 41 females (61%) with a positive perception. Ten were neutral (24.4%), 3 disagreed (7.3%), and 3 strongly disagreed (7.3%), indicating a wider range of opinions compared to males.

Overall, 19 respondents strongly agreed and 21 agreed, totaling 40 out of 61 respondents (65.6%) with a positive perception. Fourteen were neutral (23%), and 7 (11.4%) disagreed or strongly disagreed, suggesting that most respondents view the model as reducing false positives, false negatives, and false equivalence, though a significant minority were undecided or disagreed.

Table 13: Gender vs (Model is a fairer way of testing and scoring) crosstab

Gender	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Total
Male	1	1	1	3	14	20
Female	2	4	6	12	17	41
Total	3	5	7	15	31	61

Most male respondents felt that the model **is a fairer way of testing and scoring**, with 14 strongly agreeing and 3 agreeing, totaling 17 out of 20 males (85%) with a positive perception. One male was neutral (5%), and 2 (10%) disagreed or strongly disagreed, indicating strong confidence among most males.

Among females, 17 strongly agreed and 12 agreed, giving 29 out of 41 females (70.7%) with a positive perception. Six were neutral (14.6%), and 6 (14.6%) disagreed or strongly disagreed, showing that while the majority of females viewed the model as fairer, a notable portion were undecided or disagreed.

Overall, 31 respondents strongly agreed and 15 agreed, totaling 46 out of 61 respondents (75.4%) with a positive perception. Seven respondents were neutral (11.5%), and 8 (13.1%) disagreed or strongly disagreed, indicating that most respondents perceive the model as a fairer way of testing and scoring, though some expressed uncertainty or disagreement.

Table 14: Gender vs (preference for this model in my future MCQ examinations) crosstab

Gender	Strongly Disagree	Neutral	Agree	Strongly Agree	Total
Male	1	2	3	14	20
Female	5	6	11	19	41
Total	6	8	14	33	61

A large proportion of male respondents indicated that they prefer this model in their future MCQ examinations, with 14 strongly agreeing and 3 agreeing, totaling 17 out of 20 males (85%) with a positive perception. Two males were neutral (10%), and 1 (5%) strongly disagreed, showing strong overall preference among males.

Among females, 19 strongly agreed and 11 agreed, giving 30 out of 41 females (73.2%) with a positive perception. Six were neutral (14.6%), and 5 (12.2%) strongly disagreed, suggesting that while most females prefer this model, a notable portion were undecided or opposed.

Overall, 33 respondents strongly agreed and 14 agreed, totaling 47 out of 61 respondents (77%) with a positive perception. Eight respondents were neutral (13.1%), and 6 (9.8%) strongly disagreed, indicating that most respondents prefer this model for future MCQ examinations, though some remain undecided or do not prefer it.

To answer research question 3 in determining MCQ model with lower risk of misassessment of testees' true knowledge, the calculations below were done, applying epidemiological and statistical approaches:

1. Foundation Physics (PHY 101)

Experimental Event Rate (EER) – the mean in the experimental group = 51%
Control Event Rate (CER) – the mean in the control group = 42.5%

$$Absolute\ Risk\ Reduction\ (ARR) = EER - CER = 51 - 42.5 = 8.5\%$$

$$Relative\ Risk\ (RR) = \frac{CER}{EER} = \frac{42.5}{51} = 0.83$$

$$Relative\ Risk\ Reduction\ (RRR) = 1 - RR = 1 - 0.83 = 0.17 = \mathbf{17\%}$$

2. Foundation Biology (BIO 101)

Experimental Event Rate (EER) – the mean in the experimental group = 74.40%
Control Event Rate (CER) – the mean in the control group = 56.33%

$$Absolute\ Risk\ Reduction\ (ARR) = EER - CER = 74.40 - 56.33 = 18.07\%$$

$$Relative\ Risk\ (RR) = \frac{CER}{EER} = \frac{56.33}{74.40} = 0.76$$

$$Relative\ Risk\ Reduction\ (RRR) = 1 - RR = 1 - 0.76 = 0.24 = \mathbf{24\%}$$

Interpretation: In Foundation Physics and Biology, the MIMQT model reduced the risk of misassessment by 17% and 24%, respectively. In contrast, the traditional MCQ model increased the risk of mismeasuring true knowledge by the same percentages. Overall, the MIMQT model improved the accuracy of knowledge measurement for both courses. Since both ARR and RRR are high for the two foundation courses in this study, it can be said that 17% and 24% mismeasurement are statistically significant.

3.1 Ethical Considerations

Although the study was conducted using routine academic assessments, ethical considerations were observed, including informed consent. Participation was voluntary and students were assessed using standard educational instruments that posed no risk. The confidentiality of participants' information was maintained throughout the study, and all collected data were used solely for research purposes.

3.3 Limitations

A key limitation of this study is that it was conducted within a single institutional context and focused only on two foundation-level courses, which may restrict the generalizability of the findings to other disciplines or higher-level coursework. The study also relied heavily on students' self-reported perceptions, which, although valuable, are subject to response bias and may not always align perfectly with objective performance patterns. Additionally, the MIMQT model introduces a novel scoring system that some learners may initially find unfamiliar, and the study did not systematically measure the effects of prior exposure or the learning curve associated with adapting to the model. The experimental and control groups were not matched for prior academic ability, which may have influenced performance differences, even though statistical tests supported the robustness of the results.

4.0: Discussion

The results of this study demonstrate that the MIMQT model consistently outperformed the traditional MCQ

format in both Foundation Physics and Foundation Biology. Students assessed with the MIMQT model achieved significantly higher mean scores, with narrower score variation, indicating not only improved performance but also greater stability and bidirectional sensitivity in measurement. The independent samples t-test confirmed that this difference was statistically significant, supporting the conclusion that the MIMQT model measures students' true knowledge more effectively than traditional MCQs. These findings align with previous studies showing that MCQ formats that incorporate partial knowledge, or probabilistic scoring yield more accurate assessments and reduce scoring bias (Schneid et al., 2025)

The perception data further reinforces the model's accuracy and fairness. Across multiple components (including fairness in rewarding partial knowledge, reduced unfair marking, reduced misrepresentation, improved accuracy in reflecting true knowledge, and lower risk of false positives/negatives/ equivalence), the majority of respondents expressed strongly positive perceptions. Between 75% and 95% of males and 60% to 90% of females agreed or strongly agreed across components, demonstrating potential of the MIMQT model in providing a more equitable and transparent assessment. Importantly, accuracy-related components such as "confident that it reflects my true knowledge" and "the model more accurately measures my true knowledge", received more than 80% positive endorsement overall. These perceptions support the model's central purpose to more faithfully measure what learners actually know, including partial understanding that traditional MCQs fail to capture (Chang et al., 2007) (Alokozay, 2022) (Öztürk & Şahin, 2014).

Gender patterns revealed that males generally showed slightly stronger confidence, while females exhibited higher neutral responses on components related to transparency and error reduction. This pattern may reflect differences in risk perception, confidence expression, or prior exposure to alternative assessment formats (Koevoets-Beach et al., 2023) (Darabi Bazvand & Rasooli, 2022) (Madsen et al., 2013). However, this does not imply systematic bias, as studies have shown that partial credit and elimination based MCQs do not advantage one gender over another when properly implemented (Bond et al., 2013).

A notable finding is the high level of motivation associated with the MIMQT model, with over 85% of respondents indicating that the model encourages them to attempt all items. This is consistent with research showing that students feel more motivated and less anxious when assessments acknowledge partial knowledge or reduce penalties for uncertainty (Newton et al., 2025a). Likewise, the strong preference (77%) for using this model in future examinations suggests that learners not only perceive it as both fair and academically honest but also seek more sensitive alternative scoring systems for MCQs.

The risk-of-misassessment analysis provides further evidence of improved accuracy. Traditional MCQs showed a 17% mismeasurement risk in Physics and 24% in Biology, while the MIMQT model demonstrated corresponding relative risk reductions. These substantial reductions indicate that the MIMQT model more sensitively detects true knowledge and minimizes false negative scoring, supporting findings in assessment research that structured elimination or partial-knowledge scoring improves diagnostic precision (Schneid et al., n.d.). Critics of alternative MCQ formats sometimes argue that such systems introduce cognitive load or complexity, and this may explain the neutral responses observed in transparency-related components (Newton et al., 2025b). Nonetheless, these challenges can be addressed with clearer instructions and practice exposure.

Overall, the results show that the MIMQT model not only improves performance but also enhances fairness, reduces misassessment, and more accurately reflects students' true knowledge. The combination of quantitative gains, positive student perceptions, and reduced misclassification risk provides strong evidence that MIMQT is a more valid and equitable assessment approach than traditional MCQs. Future work may explore implementation at scale, effects across diverse subjects, and integration with other assessment formats for comprehensive evaluation.

Further in line with findings of this study, from time immemorial to date there are unresolved technical and policy challenges that are being inherited by today's assessment specialists (DePascale, 2025), reason for educators looking for better assessment strategies and scoring systems (Sharma, 2025). Giving zeros as an academic measurement is inequitable and produces failure rather than performance. Therefore, technical testing principles to be more relevant to the nature of classroom assessment decision making should be practiced (Ibrahim, 2023a). Ensure equitable grading practices and explore other effective grading strategies (Estayan et al., 2024), not only to remove non-academic barriers but to make grades a more accurate reflection of a student's academic performance by being bias-resistant and motivational (acknowledgement of bidirectional knowledge).

In Educational Measurement, there is no absolute zero, unlike in physical sciences where the concept of absolute zero is well conceived. For example, zero inch means absence of length, zero pound means absence of weight. But in educational measurement it is difficult to visualize a true zero in any scale used. For example, a student who scores 0 (zero) in mathematics does not imply that he/she knows nothing in mathematics (Ibrahim, 2023b). If we look at the educational measurements, you will realize that they are done at interval scale. It is therefore incorrect to score a students' performance as zero because true score is infinite, unknown, and there is no true zero point. We sometimes tell our students that 0% score on an exam does not really mean that they have no knowledge of the subject matter; rather, the exam simply did not sample the knowledge they do have (Ibrahim, 2023b). The findings show that traditional MCQs contained the knowledge which most students did not have because of the "all" or "nothing" scoring system in contrast to their counterparts in the experimental group where questions contained even the lowest knowledge possessed by testees. This means that the "0" or "1" scoring system adversely mis-assesses learners' true knowledge, turning knowledge into a dichotomous variable. Traditional MCQs format and scoring system mismeasure knowledge.

In the MIMT model, students have a 50% probability of arriving at the correct response, compared to the 25% (four-option MCQs) or 20% (five-option MCQs) probability (Sridharan & Sivaramakrishnan, 2025) afforded by traditional MCQs where distractors carry no credit. The ability to identify what does not apply within a specific subject context is a legitimate and meaningful form of knowledge, yet traditional MCQs do not recognize or reward this cognitive skill. By restricting students to only a 0.25 or 0.20 probability of earning a point, traditional MCQs effectively deny them the remaining 0.75 or 0.80 probability that reflects cognitive skill and knowledge

The MIMQT model corrects this imbalance by acknowledging that both identifying what applies and what does not apply demonstrate substantive understanding, and therefore it assigns a 0.50 probability pathway to earning credit. In this regard, withholding the larger portion of a student's opportunity to demonstrate knowledge, as traditional MCQs do, can be viewed as an unfair assessment practice that systematically disadvantages learners by reducing their legitimate chances of earning points for what they actually know. This act turns examiners into daylight academic bullies and robbers. This standpoint is in line with the findings of this study in which means, perceptions and risk calculations favor MIMQT test model.

These findings further indicate that true knowledge cannot be demonstrated solely by selecting a single best answer, as required by the Number-Right MCQ scoring approach. Instead, true knowledge is more accurately demonstrated by identifying how multiple options relate to the concept being assessed, which is the fundamental principle of the MIMQT model. Traditional NR MCQs and many other MCQ formats function, essentially as disguised True/False systems, yet they fail to award marks for the legitimate demonstration of knowledge when a student correctly identifies that an option does *not* apply to the concept in question. In contrast, the MIMQT model assesses knowledge more comprehensively by evaluating responses across several options and awarding equivalent points for correctly identifying both what applies and what does not apply. Both forms of identification reflect genuine understanding; if this were not the case, only "true" responses would earn credit while correct recognition of "false" would be dismissed as non-knowledge.

This distinction highlights that traditional MCQs evaluate knowledge based on a single answer and only in one direction, a stance that this study deems narrow and superficial because it fails to capture peripheral knowledge. The broader scope of the MIMQT model produced results that are more representative of the examinees' full spectrum of true knowledge about a given concept.

4.1: Conclusion

The findings of this study demonstrate that the MIMQT model offers a markedly more accurate, fair, and diagnostically sensitive assessment of students' true knowledge when compared with traditional MCQs. Across both Foundation Physics and Foundation Biology, students in the MIMQT group achieved significantly higher and more stable mean scores, supported by a statistically significant t-test and narrower score variability. These performance gains were reinforced by overwhelmingly positive student perceptions, particularly regarding fairness, reduced misassessment, motivation, and most critically, the model's superior accuracy in capturing true knowledge. The epidemiological risk analysis further confirmed that the traditional MCQ format carries a 17% to 24% mismeasurement risk, while the MIMQT model substantially reduces these risks, indicating a more valid and reliable measurement system.

The results also align with broader assessment literature showing that scoring approaches which acknowledge partial knowledge, reduce guessing artifacts, and incorporate probabilistic reasoning provide fairer and more meaningful evaluations of students' cognitive abilities. Conversely, the traditional "all-or-nothing" MCQ scoring method grounded in an assumed but inappropriate notion of an absolute zero in educational measurement, systematically suppresses the visibility of partial understanding and increases the likelihood of misclassification. The evidence from this study supports shifting away from dichotomous scoring systems that artificially constrain knowledge to correct/incorrect categories.

Overall, the quantitative outcomes, the strongly positive student perceptions, and the reduced misassessment risk converge to establish the MIMQT model as a more equitable, accurate, and pedagogically defensible assessment strategy. Its recognition of bidirectional knowledge not only enhances measurement validity but also promotes motivation, transparency, and fairness. These findings underscore the need for assessment specialists to adopt models such as MIMQT that align with contemporary principles of equitable grading and valid score interpretation. Future research may extend these findings by applying the model across diverse subject domains and integrating it with complementary assessment formats to further strengthen the accuracy of educational measurement.

4.2: Recommendations

1. Adopt the MIMQT model as the preferred MCQ assessment format across subjects to improve measurement accuracy and fairness.

Given the significant performance gains, reduced misassessment risk, and strong student endorsement, the MIMQT model should replace traditional MCQs in contexts where accurate measurement of true knowledge is essential. Institutions should pilot the model across multiple courses and phases of study to support wider adoption and ensure consistency in grading practices.

2. Provide structured orientation and practice opportunities to enhance transparency and reduce cognitive load associated with the new scoring system.

Although perceptions were largely positive, some respondents, especially females, expressed neutrality on components relating to clarity and error reduction. Offering demonstrations, sample items, scoring breakdowns, and low-stakes practice tests will improve students' understanding of the model and ensure that transparency issues do not hinder its effectiveness.

3. Review and reform assessment policies to eliminate unjustified "all-or-nothing" scoring and integrate partial-knowledge scoring principles into institutional standards.

Traditional MCQ scoring rests on an invalid assumption of absolute zero in educational measurement, contributing to unfair classification of learners. Institutions should revise assessment guidelines to incorporate partial-credit or elimination-based scoring models that acknowledge partial understanding, minimize false negatives/positives, and align with modern equitable grading policies. This includes training examiners on bias-resistant scoring and validation procedures.

Acknowledgements

We first and foremost thank the Almighty God for the life and chance to do this study. Heartfelt gratitude is further extended to Ms. Prisca Shula, and all authors' families for their continuous encouragement, patience and moral support during this work.

References

Adom, D., Adu-Mensah, J., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education (IJERE)*, *9*(1), 109. https://doi.org/10.11591/ijere.v9i1.20457

Alokozay, W. J. (2022). Students' perception of alternative assessment: A qualitative meta-analysis: Alternative assessment. *International Journal of Curriculum and Instruction*, *14*(2), 1419–1441.

Bond, A. E., Bodger, O., Skibinski, D. O. F., Jones, D. H., Restall, C. J., Dudley, E., & van Keulen, G. (2013). Negatively-marked MCQ assessments that reward partial knowledge do not introduce gender bias yet increase

student performance and satisfaction and reduce anxiety. *PloS One*, *8*(2), e55956. https://doi.org/10.1371/journal.pone.0055956

Chang, S.-H., Lin, P.-C., & Lin, Z.-C. (2007). Measures of Partial Knowledge and Unexpected Responses in Multiple-Choice Tests. *Journal of Educational Technology & Society*, *10*(4), 95–109.
Darabi Bazvand, A., & Rasooli, A. (2022). Students' experiences of fairness in summative assessment: A study in a higher education context. *Studies in Educational Evaluation*, *72*, 101118. https://doi.org/10.1016/j.stueduc.2021.101118

DePascale, C. (2025). *Fundamentals and Flaws of Standards-Based Testing: Lessons Learned Across Three Decades in Educational Assessment*. Taylor & Francis.

Desy, J., Harvey, A., Martin, K., Naugler, C., & McLaughlin, K. (2024). Giving partial credit during a multiple-choice question assessment reappraisal does not make the assessment process fairer. *Canadian Medical Education Journal*, *15*(2), 95–96. https://doi.org/10.36834/cmej.77957

Estayan, C., Villaver, M., & Abella, M. (2024). Exploring the Impact of Grading for Equity on Student's Academic Performance: A Qualitative Study. *Futurity Education*, *4*, 92–109. https://doi.org/10.57125/FED.2024.09.25.06

Ibrahim, A.-W. (2023a). *Reconsidering the Validity of Zero Score's Grading Practices: Imperatives of Paradigm Shift in Assessment Strategies in Nigerian Universities*. *42*, 1–10.

Ibrahim, A.-W. (2023b). *Zero Score Not Zero Knowledge: A Multi-Dimensional Assessment of Academic Knowledge and Non-Academic Behaviours* (pp. 1–570).

Kanzow, A. F., Schmidt, D., & Kanzow, P. (2023). Scoring Single-Response Multiple-Choice Items: Scoping Review and Comparison of Different Scoring Methods. *JMIR Medical Education*, *9*, e44084. https://doi.org/10.2196/44084

Koevoets-Beach, C., Julian, K., & Balabanoff, M. (2023). " I guess it was more than just my general knowledge of chemistry ": Exploring students' confidence judgments in two-tiered assessments. *Chemistry Education Research and Practice*, *24*(4), 1243–1261. https://doi.org/10.1039/D3RP00127J

Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Special Topics - Physics Education Research*, *9*(2), 020121. https://doi.org/10.1103/PhysRevSTPER.9.020121

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, *26*(8), 709–712. https://doi.org/10.1080/01421590400013495

Newton, P. M., Furby, K. H., Campbell, J., Salvi, A., Santiago, G., & Chau, M. (2025a). Negatively Marked Elimination-Format Multiple-Choice Questions Are Associated with High Cognitive Load and Poor Student Experience Compared to Single Best Answer. *Medical Science Educator*, *35*(3), 1411–1422. https://doi.org/10.1007/s40670-025-02318-7

Newton, P. M., Furby, K. H., Campbell, J., Salvi, A., Santiago, G., & Chau, M. (2025b). Negatively Marked Elimination-Format Multiple-Choice Questions Are Associated with High Cognitive Load and Poor Student Experience Compared to Single Best Answer. *Medical Science Educator*, *35*(3), 1411–1422. https://doi.org/10.1007/s40670-025-02318-7

Öztürk, Y. A., & Şahin, Ç. (2014). THE EFFECTS OF ALTERNATIVE ASSESSMENT AND EVALUATION METHODS ON ACADEMIC ACHIEVEMENT, PERSISTENCE OF LEARNING, SELF-EFFICACY PERCEPTION AND ATTITUDES. *Eğitimde Kuram ve Uygulama*, *10*(4), 1022–1046.

Parekh, P., & Bahadoor, V. (2024). The Utility of Multiple-Choice Assessment in Current Medical Education: A Critical Review. *Cureus*. https://doi.org/10.7759/cureus.59778

Schneid, S. D., Armour, C., & Brandl, K. (n.d.). Beyond right or wrong: How partial credit scoring on multiple-

choice questions improves student performance and assessment perceptions. *British Journal of Clinical Pharmacology*, *n/a*(n/a). https://doi.org/10.1002/bcp.70127

Schneid, S. D., Armour, C., & Brandl, K. (2025). Beyond right or wrong: How partial credit scoring on multiple-choice questions improves student performance and assessment perceptions. *British Journal of Clinical Pharmacology*. https://doi.org/10.1002/bcp.70127

Sharma, N. (2025, April 21). The Future of Student Assessment Beyond Traditional Methods. *Digital Engineering & Technology | Elearning Solutions | Digital Content Solutions*. https://www.hurix.com/blogs/the-future-of-student-assessment-beyond-traditional-methods/

Sridharan, K., & Sivaramakrishnan, G. (2025). Less is more? A systematic review and network meta-analysis on MCQ option numbers. *BMC Medical Education*, *25*(1), 1430. https://doi.org/10.1186/s12909-025-08026-5

Steele, S., Nayak, N., Mohamed, Y., & Panigrahi, D. (2025). The Generation and Use of Medical MCQs: A Narrative Review. *Advances in Medical Education and Practice*, *16*(null), 1331–1340. https://doi.org/10.2147/AMEP.S513119