ORIGINAL ARTICLE

TYPES OF ITEM WRITING FLAWS IN MULTIPLE CHOICE QUESTIONS IN MEDICAL EDUCATION AND THEIR EFFECT ON PASSING RATE OF STUDENTS

Sarmud Latif Awan, Shagufta Manzoor*, Irum Gillani, Ziyad Afzal Kiyani, Humza Farooq***, Sahar Khurshid†**

Department of Surgery, *Anatomy, AJK Medical College, Muzaffarabad, **Department of Public Health and Community Medicine, Health Services Academy Islamabad, ***Department of Surgery Shifa International Hospital, Islamabad, †Department of Surgery, Abbas Institute of Medical Sciences, Muzaffarabad, AJK, Pakistan

Background: The most popular written exam in health professions education is the multiple-choice style question (MCQ). This study sought to understand the effects of multiple-choice writing errors on students' academic performance in medical education. **Methods:** This descriptive study was done from December 2017 to June 2019 in Azad Jammu Kashmir (AJK) Medical College, Muzaffarabad. Ten block examinations were included. The item review committee reviewed all MCQs for flaws. Two tests from each class of MBBS were taken. The original tests containing all items were labeled as flawed tests and the result of each flawed test was evaluated. The students were graded into high, moderate and low achievement groups with scores of more than 79.9, between 50–79.9 and less than 50 percent respectively. Flawed items were then removed from the tests by the review committee and the scores of each test (standard test) were determined and compared with flawed tests and its effects were assessed in three achieving groups of students. Optical mark reading (OMR) classic-4 software was used for post-exam analysis and data were analyzed by using SPSS-25. **Results:** The passing rate ranged from 68.18% to 90.82% in flawed and 75.54% to 93.69% in standard tests. Most standard tests (7) had higher passing rates than flawed tests. **Conclusion:** Tests containing in-house developed MCQs have frequent item writing flaws and their inclusion in assessment did affect the passing rates of students but were not statistically significant.

Keywords: MCQs, flawed items, Academic achievements

Pak J Physiol 2023;19(4):49–52

INTRODUCTION

Assessment significantly impacts on students' learning and contributes to the achievement of instructional objectives. In the study of the health professions, multiple-choice questions (MCQs) are frequently employed as a form of written assessment.[1]

Concerning reliability, validity, and cost effectiveness, MCQs offer extensive material coverage for evaluating many pupils. A well-designed MCQ can evaluate several cognitive knowledge levels, ranging from memory and comprehension to application, synthesis, and analysis.[2] Moreover, MCQ examinations separate high- and low-achieving students.[3] Unfortunately, even for a well-qualified medical educator, creating a high-quality MCQ is a time-consuming, exhausting effort.[4]

There aren't many institutions in Pakistan that have medical educators who have had official training in creating MCQs. Most internal MCQs are of low quality since they are created by teachers with little to no training. There are a number of rules for creating excellent MCQs.[5] A thorough taxonomy of 31 item-writing rules has been given based on Haladyna *et al*[6], item writing principles from the National Board of Medical Examiners (NBME). The evidence-based guidelines for creating the best MCQ are frequently disregarded by item authors, which results in the creation of substandard MCQs that harm students' educational outcomes.[6]

The validity of an assessment is undermined by multiple-choice questions of poor quality. The post-examination psychometric analysis offers precise, unbiased information on the quality of the items. This quantitative analysis aids in locating various item flaws and establishes a statistical distinction between 'good' and 'poor' products.

There is no formal system for Pakistan Medical & Dental Council (PMDC) or any other supervisory entity to monitor and assess the quality of examinations. Due to shortage of medical educators and institutionalized medical education departments, local faculty members in various medical institutions are free to create MCQs in their own ways. The professor's level of expertise and experience, which differs from institution to university, has a major impact on the quality of MCQs. This study will assist in addressing the requirement for regulatory authorities to provide some system to oversee the quality aspects of MCQ-based examinations in Pakistani medical institutions.

METHODOLOGY

This non-experimental descriptive study was conducted in the AJK Medical College, Muzaffarabad from December 2017 to June 2019. This study includes 10 summative and end-of-block exams from the AJK Medical College, Muzaffarabad. These exams included two tests from 1st, 2nd, 3rd, 4th, and 5th year classes. The modules included in the study were those in which college faculty had the most influence on the development of the MCQs. The internal assessment portion accounted for 30% of the overall professional summative assessment.

Operational definitions:

1. MCQ Items:
   a) Standard Item: MCQ with no item writing flaws
   b) Flawed Item: MCQ violating one or more standard item writing principles were flagged as flawed item

It is based on Haladyna *et al*[6], item writing principles from the National Board of Medical Examiners (NBME).

2. Tests:
   a) Standard test: Test after exclusion of flawed items
   b) Flawed test: Test inclusive of flawed items
3. Groups of students based on scores in tests:
   a) High achievers: Those who score 80% or above marks
   b) Moderate achievers: Those who score between 50–79.9% marks
   c) Low achievers: Those who score less than 50% marks

The Item Review Committee of Department of Examination, AJK Medical College examined every MCQ for errors in item authoring. Examinations from first-, second-, third-, fourth-, and fifth-year classes yielded two tests from each category. The study included tests from summative and end-of-block assessments, tests with post-examination statistical data including the reliability of the test, difficulty index, point biserial and discrimination indices of items. Number of students was 90 or more per test, number of MCQs was 50 or more per test, and MCQ items were written by local faculty. With all test items included, the first result of each test (flawed test) was produced, and students were classified appropriately into high, moderate, and low groups.

RESULTS

There were 145 (29%) flawed items in 500 MCQ items in 10 tests. In these flawed items, the five most common ones were K-type (26%), negative stem (20%), non-homogenous distracters (17%), all of the above (14%) and implausible distracters (9.6%). These flaws accounted for 86.6% of all flaws (Table-1, 2).

Table-1: Frequency of flaw items

Test No	Number of students	Total MCQ items	Number of flawed items
1	106	50	13
2	98	50	7
3	88	50	16
4	95	50	12
5	93	50	18
6	95	50	17
7	95	50	19
8	106	50	12
9	87	50	15
10	87	50	16
Total	**950**	**500**	**145**

Table-2: Types of flawed items

Type of flaws	No of Flaws
Negative stem	29
K-type	38
Implausible distracters	14
Unfocussed stem	9
Unequal length of distracters	3
None of above	0
Logical cues	0
More than one flaws	2
True-False	0
Trival Content	2
All of above	20
Repeat words/grammatical errors	0
Complex partial type	3
Non homogenous distracters	25
Total	145

There were observed differences in the passing rates of students in flawed and standard tests as shown in Table-3. The passing rate ranged from 68.18% to 90.82% in flawed and 75.54% to 93.69% in standard tests. In tests 1, 5 and 10 the student's pass percentage was higher in flawed tests. In remaining tests, standard tests had a higher pass percentage. One-hundred-sixty-three students failed in flawed tests while 119 students failed in standard tests. Hence 44 more students could get through examinations if the tests had no flawed items. Mann Whitney U test was used to determine the statistical significance in the passing rates of flawed and standard tests (Table-4, 5). In this study Null hypothesis could not be rejected as in Mann-Whitney U test the 2-tailed significance was 0.226.

Table-3: Passing rates (%) of students in flawed and standard tests

Test No.	Flawed	Standard
1	81.13	78.24
2	90.82	91.83
3	68.18	75.54
4	88.42	93.68
5	88.17	83.87
6	81.05	93.69
7	85.26	90.52
8	77.36	84.90
9	83.90	90.65
10	83.86	81.60

Table-4: Statistical analysis of passing rates in flawed and standard tests

Ranks				
	Groups	n	Mean Rank	Sum of Ranks
Passing rates	Flawed	10	8.90	89.00
	Standard	10	12.10	121.00
	Total	20		

Table-5: Test statistics Man-Whitney-U

Test Statistics[a]	
	Passing rates
Mann-Whitney U	34.000
Wilcoxon W	89.000
Z	-1.209
Asymp. Significance (2-tailed)	0.226
Exact Significance [2*(1-tailed Sig.)]	0.247[b]

DISCUSSION

These tests had high rate of flawed items ranging from 14% to 38% (mean 29%). The five most common flaws were K-type (26%), negative stem (20%), non-homogenous distracters (17%), all of the above (14%), and implausible distracters (9.6%). Out of all flawed items, these flaws accounted for 86.6%. These findings are very similar to the findings of Downing[8] where the most common five flaws accounting for 90% of all flaws were an unfocused stem, a negative stem, all of the above, none of the above options and partial K-type items in his study. Tarrant and James[9] had similar findings in their study, where the most common eight flaws encompassed 85% of all flaws were negative stem, unnecessary information in the stem, no correct or more than one correct answer, implausible distracters, greater detail in correct option, logical clues and word repeats. The item writing flaws found in this study were the ones which are well-reported in medical literature.

Training and experience of the faculty involved in item writing directly determine the quality of MCQs items. In AJK Medical College, though the faculty is highly trained in their related subjects, they have little training in assessment methodology. There are few such training opportunities in our country with a limited number of formally trained medical educationists. Furthermore, the regulatory authorities do not impose any requirement for such training. In order to write high-quality MCQs, one must not only be aware of item writing principles but also have supervised training. Only through training and experience the faculty develops the ability to write high-quality MCQs. These item writing flaws can be corrected by faculty development programs in medical institutions. In the presence of item writing flaws, the test results validity is threatened by construct irrelevance variance. According to Downing the training of the faculty for item writing and pre-examination item review for correction of these flaws improves validity of test results.[1]

This study revealed that in tests 2, 3, 4, 6, 7, 8, 9 the pass percentage of students was higher in standard tests than in flawed tests. However, in tests 1, 5, and 10, the pass percentage was less in standard tests than in flawed tests. In most of the tests more students passed when flawed items were excluded from the tests. Thus, the flawed items resulted in higher failure rates in flawed tests and acted as a disadvantage for these students. These results were similar to the results of two different studies by Downing.[7,8]

In three examinations (tests 1, 5 and 10), the pass percentage was higher in flawed tests than standard tests. These results were similar to the results found in the study of Tarrant where flawed items had a positive effect on borderline students and their inclusion resulted in the passing of greater number of students.[9] Tests 2 and 10 had small differences in the pass percentage of students in standard and flawed tests similar to the results found in the study of Wadi.[10]

Including flawed items in the test leads to construct-irrelevant errors in the tests.[7] Therefore, assessment does not determine the true competence level of the students and lack construct validity. These inaccuracies resulting from the inclusion of flawed items in tests lead to the failure of the students who deserve to pass and the passing of students who deserve to fail.

Przymuszała *et al*[11] found that guidelines on writing multiple choice questions were a well-received and effective faculty development intervention. Most of the medical teachers in undergraduate institutions excel in their respective specialized fields but have little insight into the complexities of the assessment in health professional education. At the same time assessment has become a specialized field even for medical educationists. The quality of assessment cannot be improved without educating 'specialty-trained' teachers in assessment methods.

Fayyaz Khan *et al*[12] examined multiple choice questions from 2009 to 2011, finding that technical flaws in the questions were common, particularly in test-wiseness and irrelevant difficulty, highlighting the need for better MCQ quality.[12]

CONCLUSION

Tests containing in-house developed MCQs have frequent item writing flaws. The use of flawed items in the assessment has various unexpected negative consequences on students' academic achievements. In spite of acceptable psychometrics of flawed items in the tests, their inclusion in assessment did affect the passing rates of students but it was not statistically significant.

RECOMMENDATIONS

- This was a short study, a step forward but undoubtedly insufficient to settle all disagreements. A bigger, ideally multicentre, randomized control research will be required to address this problem.

- Programs for faculty development may offer the foundation needed to improve the standard of evaluation in medical institutions. The faculty frequently cites a select number of writing errors as particularly prevalent. The institution particularly the medical education department is accountable for recognizing and fixing these persistent problems during faculty development.
- The regulatory bodies must support and enable medical education departments and educators to take on the role of guardians of quality evaluation in medical institutions.

LIMITATIONS

This sample does not accurately represent the best type of multiple-choice questions used during assessments at other medical colleges nationwide because the study was only conducted at one public medical institution.

The study was also impacted by the training of the item writers because they varied in their levels of expertise across subjects. Only a small number of faculty members who had attended item writing workshops had received training. These professors were not equally spread across subjects, with some having higher quality item writers than others.

REFERENCES

1. Downing SM, Yudkowsky R, (Eds). Assessment in health professions education. New York: Routledge; 2009.p.149–84.
2. Palmer, EJ, Devitt, PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. BMC Med Educ 2007;7:49.
3. Schuwirth LW, van der Vleuten, CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004;38(9):974–9.
4. DiBattista D, Kurzawa, L. Examination of the quality of multiple-choice items on classroom tests. Can J Scholarsh Teach Learn 2011;2(2):4.
5. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. National Board of Medical Examiners: Philadelphia, PA; 1998.
6. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ 2002;15(3):309–34.
7. Downing SM. Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? Acad Med 2002;77(10 Suppl):S103–4.
8. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract 2005;10:133–43.
9. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 2008;42(2):198–206.
10. Wadi M, Rahim AFA, Yusoff MSB, Baharuddin, KA. The effect of MCQ vetting on students' examination performance. Educ Med J 2014;6(2):e16–26.
11. Przymuszała P, Piotrowska K, Lipski D, Marciniak R, Cerbin-Koczorowska M. Guidelines on writing multiple choice questions: A well-received and effective faculty development intervention. SAGE Open 2020;10(3):2158244020947432.
12. Fayyaz Khan H, Farooq Danish K, Saeed Awan A, Anwar M. Identification of technical item flaws leads to improvement of the quality of single best Multiple Choice Questions. Pak J Med Sci 2013;29(3):715–8.

Address for Correspondence:

Dr Sarmud Latif Awan, Department of Surgery, Abbas Institute of Medical Sciences, Muzaffarabad, Azad Kashmir, Pakistan. **Cell:** +92-334-5285882

Email: sarmadawan786@hotmail,com

Received: 26 Sep 2023 **Reviewed:** 7 Nov 2023 **Accepted:** 7 Nov 2023

Contribution of Authors:

SLA: Concept, design of work, article writing
SM: Data analysis
IG: Data analysis, maintain interpretation
ZAK: Critical review
HF: Data collection
SK: Data collection

Conflict of Interest: None
Funding: None