ORIGINAL ARTICLE

EFFECT OF ITEM WRITING FLAWS IN MULTIPLE CHOICE QUESTIONS ON STUDENTS' ACHIEVEMENT GROUPS AND RELIABILITIES OF TESTS IN AJK MEDICAL COLLEGE

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

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

Background: Multiple-choice question (MCQ) is a commonly used tool for written assessment. Purpose of this study was to determine the effect of item writing flaws in MCQs on students' academic achievements and test reliabilities. **Methods:** This study was conducted at AJK Medical College Muzaffarabad from Dec 2017 to Jun 2019. Ten summative tests were included. The item review committee examined every MCQ for errors in item writing. The initial tests with all items were deemed to be flawed tests. The outcomes of these tests were assessed and students were ranked into three achievement groups based on their scores, i.e., high, moderate, and low achievers with scores of >79.9, 50–79.9, and <50% respectively. Once the review board eliminated the flawed items, the scores of each test (the standard test) were calculated, compared to the flawed tests and its impacts were evaluated between three achievement groups. The post-exam analysis was done using the optical mark reading classic-4 programme. Data were analyzed using SPSS-25. **Results:** In only test No. 6 the null hypothesis, i.e., there is no effect of flawed items on students' academic achievements was statistically rejected (p<0.05). Among high, moderate and low achievers between flawed and standard tests, moderate achievers and low achievers had statistically significant correlation (p<0.003 and <0.044 respectively). The flawed tests had better reliabilities than standard tests with statistically significant difference (p<0.012**) Conclusion:** Flawed items negatively affect high and moderate achievers and affect low achievers positively. Flawed tests had better reliabilities.

Keywords: Academic achievements, flawed items, MCQs

Pak J Physiol 2024;20(1):41–4

INTRODUCTION

Students' learning is greatly influenced by assessment which also helps to accomplish curricular goals. Multiple-choice question (MCQ) is a commonly used tool for written assessment in health professions education.[1] Properly constructed MCQs can test different levels of cognitive knowledge from recall, comprehension to application, synthesis and analysis.[2] MCQs tests also discriminate between high and low achieving students.[3] However, construction of a high quality MCQ is a time consuming, laborious and taxing task, even for a properly trained medical educationist.[4]

Few institutions in Pakistan have properly trained medical educationists with formal training in writing MCQs. A large number of in house MCQs are developed by faculty with little or no training; hence, these are mostly of low quality. There are several guidelines for construction of high quality MCQs.[5] A detailed taxonomy of 31 item-writing rules has been described.[6] The evidence-based recommendations for the construction of one best MCQ are often violated by item writers which leads to the production of flawed MCQs items with adverse effects on student's academic achievements.[6]

Medical education is being supervised by PMDC in Pakistan. But there is no organized mechanism for supervision and evaluation of standards of examinations by PMDC or any other supervisory authority in Pakistan. In view of scarcity of medical educationists and established medical education departments, local faculty members in different medical institutions are at liberty to develop MCQs in their own way. The quality of MCQs is primarily dependent on the experience and training of faculty, which varies from institution to institution.

This study will help in determining the effect of item writing flaws in multiple choice questions in basic and clinical sciences on students' academic achievements and on test reliabilities rectifying the need for creation of some mechanism by regulatory authorities to supervise the quality aspects of MCQs based examinations in medical institutions of Pakistan.

METHODOLOGY

This study was a non-experimental descriptive study carried out from Dec 2017 to Jun 2019 in AJK Medical College, Muzaffarabad. This study was approved by the ethical committee of AJK Medical College. Ten summative and end-of-block assessments

in AJK Medical College were included in this study. These tests were taken from assessments of 1st, 2nd, 3rd, 4th and 5th year classes with two tests from each class. Modules included in study were those in which college faculty had maximum input in terms of MCQ construction and these were the part of internal assessment that constituted 30% of final professional summative assessment.

All MCQs were reviewed by the item review committee for item writing flaws in the examination department of AJK Medical College. Tests from summative end of block assessment with post-examination analysis statistical data interms of reliability of test, difficulty index, point biserial and discrimination indices of items, with 90 or more student per test, 50 or more number of MCQs written by local faculty per test, with reliability of 0.6 or greater were included in the study. The initial result of each test (flawed test with all items in the test) was obtained and students grouped accordingly into high, moderate and low achieving groups. Each test item was reviewed by review committee. The review committee comprised of one writing expert and one relevant subject specialist.

There was only one item writing expert who was the permanent member of each review committee. There were different subject specialists for each MCQ paper from different disciplines. Flawed items were then removed from the tests by review committee.

The scores of each test (standard test without flaw items) were then determined and students were graded into high, moderate and low achievers. The scores of each flawed test and standard test were compared and their effects were determined in three achieving groups. Optical mark reading (OMR) classic-4 software was used for post-exam analysis of flawed items.

RESULTS

The observed differences in the number of students in different achievement groups in flawed and standard tests in this study are shown in Table-1. In one test (tests No. 6) the null hypothesis, i.e., there was no statistical significant association of flawed items to achievement groups was rejected with 95% confidence interval. In all other tests (Test 1, 2, 3, 5, 7, 8, 9, 10) null hypothesis could not be rejected on statistical basis. These results showed that there was significant association of presence or absence of flawed items to achievement groups in one test. There was no statistical significant association of flawed items to achievement groups in nine tests in this study. (Table-1).

Cumulative differences were observed in each achievement group of students in all tests. These observed differences are summarized in Table-2.

Table-1: Differences in the frequency of achievement groups in flawed and standard test

Achievement Groups	Pass	Fail	High	Moderate	Low	*p*
Test-1						
Flaw	86	20	0	86	20	0.116
Standard	84	22	4	80	22	
Test-2						
Flaw	89	9	2	87	9	0.09
Standard	90	8	9	81	8	
Test-3						
Flaw	60	28	2	58	28	0.474
Standard	67	21	3	64	21	
Test-4						
Flaw	84	11	4	80	11	0.432
Standard	89	6	5	84	6	
Test-5						
Flaw	82	11	2	80	11	0.46
Standard	78	15	4	75	15	
Test-6						
Flaw	77	18	7	70	18	0.029
Standard	89	6	10	79	6	
Test-7						
Flaw	81	14	5	76	14	0.406
Standard	86	9	8	78	9	
Test-8						
Flaw	82	24	2	80	24	0.219
Standard	90	16	5	85	16	
Test-9						
Flaw	73	14	6	67	14	0.667
Standard	77	10	7	70	10	
Test-10						
Flaw	73	14	9	64	14	0.346
Standard	71	16	4	67	16	

Table-2: Number of students in achievement groups in flawed and standard test

Achievement group	Flawed tests	Standard tests
High achievement group	39	59
Moderate achievement group	748	763
Low Achievement Group	163	129

Cumulative differences of achievement groups (high, moderate, and low)

The high achievers in flawed tests were 39, and in standard tests they were 59. There was a difference of 20 students. Inclusion of flawed items in these tests negatively affected scores of high achievers. There were increase number of high achievers in standard tests than flawed tests but the correlation was not statistically significant (p>0.05). The moderate achievers in flawed tests were 748 and in standard tests there were 763. There was a difference of 15 students. Inclusion of flawed items in these tests negatively affected scores of moderate achievers.

There were increase number of moderate achievers in standard tests than flawed tests and the correlation was statistically significant (p=0.003). The low achievers in flawed tests were 163 and in standard tests there were 129. There was a difference of 34 students. Inclusion of flawed items in these tests negatively affected scores of low achievers. There was decrease in number of low achievers in standard tests than flawed tests and the correlation was statistically significant (p=0.044) (Table-3).

Table 3: Correlation of achievement groups between flawed and standard tests

		High achievement students in standard test	High achievement students in flawed test
High achievement students in standard test	Pearson Correlation	1	0.325
	Sig. (2-tailed)		0.360
	N	10	10
High achievement students in flawed test	Pearson Correlation	0.325	1
	Sig. (2-tailed)	0.360	
	N	10	10
		Moderate achievement students in standard test	Moderate achievement students in flawed test
Moderate achievement students in standard test	Pearson Correlation	1	0.829**
	Sig. (2-tailed)		0.003
	N	10	10
Moderate achievement students in flawed test	Pearson Correlation	0.829*	1
	Sig. (2-tailed)	0.003	
	N	10	10
		Low achievement students in standard test	Low achievement students in flawed test
Low achievement students in standard test	Pearson Correlation	1	0.645*
	Sig. (2-tailed)		0.044
	N	10	10
Low achievement students in flawed test	Pearson Correlation	0.645**.	1
	Sig. (2-tailed)	0.044	
	N	10	10

*Correlation is significant at the 0.01 level (2-tailed), **Correlation is significant at the 0.05 level (2-tailed).

The null hypothesis could not be rejected for high achievement group in this study. There were observable difference in high achievement group in flawed and standard test but this difference was not statistically significant. There was statistically significant association of flawed items in moderate and low achievements groups. The null hypothesis was rejected in these groups with p=0.003 and 0.044 respectively. Kuder-Richchardson-20 Formula (KR-20) was used to determine the reliabilities of flawed and standard tests. The reliabilities ranged from 0.6 to 0.78. The mean of reliabilities for flawed tests and standard tests were 0.72 and 0.65 respectively. In all tests reliabilities of flawed tests were better than standard tests. (Table-4).

Table-4: Reliabilities of flawed and standard tests

Reliability of Tests	Kuder-Richardson Formula 20
Test-1	
Flawed	0.66
Standard	0.60
Test-2	
Flawed	0.74
Standard	0.72
Test-3	
Flawed	0.68
Standard	0.62
Test-4	
Flawed	0.78
Standard	0.72
Test-5	
Flawed	0.69
Standard	0.63
Test-6	
Flawed	0.78
Standard	0.66
Test-7	
Flawed	0.78
Standard	0.66
Test-8	
Flawed	0.68
Standard	0.62
Test-9	
Flawed	0.66
Standard	0.60
Test-10	
Flawed	0.76
Standard	0.71

Mann Whitney U test was used to determine statistical significance between reliabilities of flawed and standard test. The difference in reliabilities of flawed and standard tests was found to be statistically significant (p=0.012) in the current study, so null hypothesis, i.e., there is no significant difference in reliabilities of flawed and standard test, could be rejected. (Table-5).

Table-5: Man-Whitney test statistics

Test parameters	Reliabilities
Mann-Whitney U	17.000
Wilcoxon W	72.000
Z	-2.512
Asymp. Sig. (2-tailed)	0.012
Exact Sig. [2*(1-tailed Sig.)]	0.011[a]

[a]Not corrected for ties

DISCUSSION

In this study, the students were grouped into high, moderate, and low achievement group on the basis of their performance. The presence of flawed items in the tests had a negative effect on the results. Exclusion of flawed items led to increase in number of students from in high and moderate achievement groups, and decrease in number of students in low achievement group. These findings are similar to studies by Downing[7,8] and Tarrant[9]. Inclusion of flawed items in the test not only contorted the pass/fail decisions but also negatively affected the process of awarding grades to the students in the test. An observable number of students could not achieve >80% score because of the flaw items in the test. Similarly, 34 students who deserved to be in moderate achievement group fell in low achievement group.

The prime objective of the assessment is not to award grades to the students but to differentiate between high and low achieving students. In our study, the actual boundaries of the three achievement groups of students were distorted by these flawed items, as high performing students gave the impression of being moderately

performing students and moderately performing students as low performing students. Inclusion of flawed items in the tests greatly compromised the authenticity of grading decision in the assessment.

According to Axelson *et al*[10], the reliability of a test is an estimate of proportionate amount of random error in the data. Reliabilities of all tests (standard) in this study got decreased when flawed items were removed from the test. There was a decrease in the reliability of 10 standard tests after removal of flawed items from the tests. These differences were statistically significant in this study. In a study by Downing[7], there was no difference in the reliability of flawed and standard scales. Tarrant *et al*[9] used KR-20 for measurement of internal consistency of 10 tests. The KR-20 ranged from 0.54 to 0.87. The reliability estimates in our study were similar to findings of KR-20 reliability. The reliability of 8 (out of 10) standard tests in Tarrant[9] study was lower than the reliability of total (flawed) scale even after correction for the length of tests. We also had similar findings where reliability of 10 standard tests was lower than the reliability of flawed tests. Two important determinants of the reliability of a written test are the length of test and performance of items on test.[10] When these flawed items were removed from the test, the length of the test got reduced and as a result the reliabilities of these tests (standard) also decreased. There was an observable reduction of reliability where maximum and minimum items were removed from the test. The reliability of tests was more influenced by the length and number of items in a test. The reliability of standard tests reduced after removal of flawed items with acceptable psychometrics.

CONCLUSION

The use of flawed items in the assessment results in negatively affecting high and moderate achievers and positively affecting low achievers. Inclusion of flawed items in tests greatly compromised the authenticity of grading decision in the assessment. Overall reliabilities of flawed test were greater than standard tests.

RECOMMENDATIONS

This was a small study, a way forward but certainly not enough to resolve all controversies. A larger, preferably multi-centre, randomized control study will be required to resolve the issue. Faculty development programmes can provide the platform for raising the quality of assessment in medical institutions. It is the responsibility of institution especially Medical Education Department to identify and rectify these commonly repeated flaws during faculty training.

REFERENCES

1. Downing SM, Written tests. In: Downing SM, Yudkowsky R. (Eds). Assessment in health professions education. New York: Routledge; 2009.pp 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. Philadelphia: National Board of Medical Examiners; 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):309–33.
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(2):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. Axelson RD, Kreiter CD. Reliability. In: Downing SM, Yudkowsky R. (Eds). Assessment in health professions education. New York: Routledge; 2009.pp 57–73.

Address for Correspondence:

Dr. Sarmad Latif Awan, Associate Professor, Department of Surgery, Abbas Institute of Medical Sciences/AJK Medical College, Muzaffarabad. **Cell:** +92-334-5285882
Email: sarmadawan@hotmail.com

Received: 26 Sep 2023 **Reviewed:** 4 Jan 2024 **Accepted:** 15 Feb 2024

Contribution of Authors:
SLA: Concept, design of work, article writing
SM: Data analysis
IG: Data analysis
ZAK: Critical review
SK: Data collection
HF: Data Collection

Conflict of Interest: None
Funding: None