Int.j.pathol.2019;17(2)84-88

Original Article

Impact of Flawed Multiple-Choice Questions on Tests and Student Achievements in Medical Education

Abdul Khalid Awan
Department of Medicine and Medical Education
Azad Jammu & Kashmir Medical College, Muzaffarabad, Azad Kashmir

ABSTRACT
Objectives: The objectives were to determine the frequency of flawed items in MCQ based tests in Azad Jammu and Kashmir Medical College (AJK-MC) Muzaffarabad, to compare the Quality of examination before and after removing the flawed MCQ items from the tests, to compare student's achievements on standard (without flawed items) and flawed tests and to find out the student group/ groups whom result is maximally affected by the presence of flawed MCQs items in examinations.
Methods: 10 summative examinations in AJK-MC Muzaffarabad were included in the study. The frequencies of flawed MCQ items with different types of item writing flaws in tests were identified. The student scores in each test were compared before and after removal of flawed items and its effects were evaluated in high, moderate and low achieving groups of students. The Mann-Whitney U test was used for comparison of student's scores on standard and flawed tests. The internal consistency reliability of tests was measured by Kuder-Richardson (KR-20) formula. Chi Square and Fischer Exact tests were used to determine the association between different categorical variables in the study.
Results: There was high frequency of flawed items ranging from 7 to 52 percent in different tests. There were seven most common item writing flaws. More students failed in flawed as compared with standard tests in most of examinations.
Conclusion: Item writing flaws are common in in-house developed MCQ tests and there are unintended negative consequences of using flawed items on test scores and student achievements.
Keywords: Flawed MCQs, Item writing flaws, Standard and Flawed tests

Background

Multiple-choice question (MCQs) is the most useful and commonly used format for written assessments in health professional education.[1]This format allows wide content coverage for assessment of large number of candidates with high reliability, validity and cost effectiveness. Assessment has a powerful influence on students learning and properly constructed MCQs can test different levels of knowledge in cognitive domain from recall, comprehension to application, synthesis and analysis.[2] Good quality MCQs tests can also discriminate between high and low achieving students.[3] However, construction of a high quality MCQ is time consuming, laborious and taxing task, even for a properly trained medical educationist.[4] It needs supervised training as there are several principles for item writing and violation of these guidelines lead to the production of flawed items.[5,6] Students are at risk of being tested with these flawed MCQs for their competency and academic performance, which is detrimental to their success.

The study was designed to evaluate the frequency of Flawed MCQ items and their impact on student achievement in high stake examinations of Azad Jammu and Kashmir Medical College (AJK-MC). The effects of flawed MCQs on different achievement groups of students were analyzed and most common item writing flaws in MCQ based tests were also evaluated.

AUTHOR CORRESPONDENCE:
Dr. Abdul Khalid Awan
Professor of Medicine and Co-Associate Dean
Medical Education, AJK Medical College, Muzaffarabad
E-mail:khalid_awanajk@yahoo.com

Methods

A psychometric, non-experimental study was performed in the department of medical education in Azad Jammu and Kashmir Medical College (AJK-MC) Muzaffarabad. Three hypotheses were postulated for the study; a-there will be significant differences in

pass/fail rates of students in flawed and standard tests, b-there will be significant differences in high, moderate and low achievement groups of students in flawed and standard tests, c-there will be significant differences in the reliabilities of flawed and standard tests. The operational definitions used in the study were:

Standard test: Test after exclusion of flawed items. **Flawed test:** Test inclusive of flawed items. The achievement groups of students were defined as:

High achievers: those who scored 80% or above.

Moderate achievers: those who scored between 50%-79.9% and **Low achievers:** those who scored less than 50% marks. Ten summative MCQ tests (table-2) were examined which were administered to medical students at the end of the module or end of Block examinations in AJK-MC Muzaffarabad. The study was approved by institutional review board of AJK-MC. The tests were taken from assessment of 1st to 4th year classes and examinations were labeled as Test-1 to Test-10 in chronological order of data collection. The tests included both pre-clinical and clinical subjects. Post examination analysis of all items was performed by using OMR classic-4 software. The first result of each (flawed) test was compiled with all items in the test and students then ranked ordered accordingly in high, moderate and low achieving groups.

A total number of 685 MCQs items were included in the study. The item review committee (one item writing expert and one relevant subject specialist) identified the number of flawed items in the test. Then a second result of the test was compiled after removing flawed items from the (standard) test. The students were again ranked ordered in high, moderate and low achieving groups. The student's scores on each test were compared before and after removal of flawed items and its effects were evaluated in high, moderate and low achieving groups. The item review committee also identified frequency and different types of item writing flaws in MCQs.

The Mann-Whitney U (a non-parametric) test was used for comparison of student's scores on standard and flawed tests. The internal consistency reliability of tests was measured by Kuder-Richardson (KR-20) formula. Chi Square and Fischer Exact tests were used to determine the association between different categorical variables in this study.

Results

A total number of 986 students appeared in ten examinations. The total numbers of MCQs were 685, with a range of 50-100 items per test. The proportion of flawed items in these tests ranged from 7-52 %. The total numbers of flawed items were 152 (Table-1). Among these flawed items, the six most common flaws were negative stem (25%), implausible distractors (19%), unfocussed stem (12.5%), unequal length of distractors (7.8%), none of above (6.6%) and all of above (6.6%).These flaws were accounting for 78% of all flaws. The negative stem was the most common item writing flaw followed by implausible distractors and unfocused stems. There were only a small proportion of other flaws like grammatical errors, non-homogenous distractors and repeat words. There was more than one flaw in 11 (7%) items.

Table-1

Type of flaws	No of Flaws
Negative stem	38
Implausible distractors	29
Unfocussed stem	19
Unequal length of distractors	12
None of above	10
Logical cues	9
More than one flaws	11
True-False	8
All of above	10
Repeat words/grammatical errors	3
Complex partial type	2
Non homogenous distractors	1
Total	152

There were observed differences in the passing rates of students in flawed and standard tests. The passing rates ranged from 45% to 96% in flawed and 63% to 99% in standard tests. The passing rate of students was higher in flawed tests 2 and 8. The passing rate was equal in flawed and standard tests in test-4. In all other tests passing rate was higher in standard tests as more students were failing in flawed tests. A total number of 184 students failed in flawed tests while only 156 students failed in standard tests. There was a discordance of 28 students who would have passed the examinations had the flawed items were removed from the tests. The passing rates of students in flawed and standard tests were analyzed by Mann Whitney U test for statistical significance. The 2-tailed significance of Mann-Whitney U test was 0.59, so null hypothesis could not be rejected in the study. Although there were differences in passing rates of students in majority of tests, but these differences never reached statistical significance in the study.

A wide range of cumulative differences observed in high, moderate and low achievement group of students in 10 tests as summarized in tables 2 and 3. The Chi Square and Fisher exact tests were used to

analyze the association of presence or absence of flawed items with differences in achievements groups (High, Moderate and Low) in the tests. The number of students who achieved more than 80% scores (high achievers) in flawed tests was 9 while 17 students scored more than 80% in standard tests. The discordance occurred in 8 students. There was negative effect on scores of high achieving students with inclusion of flawed items in these tests. Although there were more students in high achievements groups in standard tests as compared with flawed tests the correlation however, was not statistically significant ($p < 0.07$).The number of students who achieved scores between 50 to 79.99 % (moderate achievers) was also different in flawed and standard tests. There were 793 students in this category in flawed tests while 813 students scored moderately in standard test. The correlation in groups of moderate achieving students in flawed and standard tests was statistically significant ($p < 0.002$).There were 184 students who scored less than 50% (low achievers) in flawed tests while 156 scored less than 50% in standard tests. The correlation in groups of low achieving students in flawed and standard tests was also statistically significant in the study ($p < 0.001$).

Table-2: Distribution of study participants in different assessment groups and description of various modules assessed in current study

Achievement Group	Number of students in Flawed tests	Number of students in standard tests
High Achievement Group	9	17
Moderate Achievement Group	793	813
Low Achievement Group	184	156

Tests included in the study

Test Number and Name of Module	Number of MCQs	Number of Students
1-Cell and Molecular Biology-1st year	50	93
2-Cardiovascular Module-1st year	100	99
3-Gastro-intestinal Module-2nd year	100	99
4-Endocrine, Metabolism and Reproduction-2nd year	60	97
5-Endocrine, Metabolism and Reproduction-3rd year	100	99
6-Inflammation, healing and Immunity-3rd year	50	101
7-Blood and Immunity-1st year	50	103
8-Cardiovascular Module-2nd year	75	99
9-Respiratory and CVS Module-3rd year	100	95
10-Legal Module-4th year	50	102

The reliabilities of flawed and standard tests were measured by Kuder-Richchardson-20 Formula (KR-20). The reliabilities of flawed tests were better than the reliabilities of standard test in most examinations. The range of reliability was from 0.5 to 0.8 in different tests. The mean reliability of the flawed tests was better than the mean reliability of standard tests (0.7 and 0.65 respectively). The means of reliabilities of tests were compared for significance by using Mann Whitney U test. There was no statistically significant differences in the reliabilities of flawed and standard tests ($p<0.34$), despite observed differences null hypothesis could not be rejected.

Discussion

Several aspects of flawed items and impact of their presence on tests and student achievements were analyzed in the study. The passing rate of students was high on standard test in seven examinations (Tests 1, 3, 5, 6, 7, 9, and 10). The passing rate was low on standard tests in three examinations (Tests 2, 4 and 8). In majority of tests more students passed while flawed items were removed from the tests. The presence of flawed items was contributing towards higher failure rates in flawed tests and served as disadvantage for most students. Similar results were also found in two different studies by Steven M Downing.[7, 8]

In three examinations more students passed on flawed tests as compare with standard tests. The findings in these tests were similar to findings of Marie Tarrant where flawed items were found more beneficial for borderline students as more students passed when these items were included in the tests [9].In four tests in this study there was very small difference in passing rates on flawed and standard tests (Tests 1, 3,4 and 10). Wadi M found same results in an experimental study where he compared two groups of students in mock examinations[10]. The presence of flawed items introduced construct-irrelevant error in the tests; hence assessment did not represent the true level of competence of the students and also lacked construct-validity evidence for assessment [7]. As a result of these inaccuracies due to the presence of flawed items, students who deserved to pass an examination were failed and those who deserved to fail were passed and promoted to next stage in their academics. This type of inaccuracy and lack of quality in assessment adversely affects the morale and future career of students. It also poses serious questions about the legitimacy and integrity of examination process and quality.

The analysis of previous studies had shown different effects on passing rates in flawed and standard tests. In Downing studies[11] flawed items were associated with higher failure rates, Tarrant [12] found higher passing rates with flawed items and in Wadi's study[13] there was no difference in pass-fail rates of students either tested with vetted (standard) or flawed (non-vetted) MCQs.

Table-3 Comparison of results before and after removing the flawed MCQs from the tests

Achievement groups	Tests																			
	Test-1		Test-2		Test-3		Test-4		Test-5		Test-6		Test-7		Test-8		Test-9		Test-10	
	flawed	standard	flawed	standard	flawed	standard	flawed	standard	flawed	standard	flawed	standard	flawed	Standard	flawed	standard	flawed	standard	flawed	standard
Total Pass	89	90	82	78	88	89	94	93	95	98	46	61	86	93	78	71	64	67	80	90
Total Fail	9	8	11	15	11	10	3	4	4	1	55	40	17	10	21	28	31	28	22	12
High ach*	2	9	2	4	1	1	2	0	1	1	0	0	1	1	0	0	0	0	0	1
Mod Ach*	87	81	80	74	87	88	92	93	94	97	46	61	85	92	78	71	64	67	80	89
Low Ach*	9	8	11	15	11	10	3	4	4	1	55	40	17	10	21	28	31	28	22	12

*High ach, Mod ach, Low ach Number of high, moderate and low achieving students respectively

In this study there were also variable effects of flawed items on passing rates of students. There was no uniform pattern of influence of flawed items in all examinations. However, it was obvious, as it was from the results of majority of previous studies that presence of flawed items distorted pass fail decisions and assessment did not reflect the true level of competence of examinees. The presence of flawed items in this study was also associated with significant disadvantage for students (though if failed to achieve the statistical significance) as more students failed in majority of flawed tests due to the presence of flawed items.

There were three achievement groups in the study based on performance in tests. The results of all three groups of students were negatively affected by the presence of flawed items in the tests. The number of students almost doubled (from 9 to 17) in high achievement groups (students scoring more than 80%) when flawed items were excluded from the tests. There were only 793 students in moderate achievement group in flawed tests as compared with 813 students in standard tests. Similarly in low achievement group 184 students failed in flawed tests while only 156 students failed in standard tests. These results were in accordance to the results of previous findings in studies by Steven M Downing and Marie Tarrant. It was not only the pass/fail decisions which were distorted, but the whole process of awarding grades was affected by the presence of flawed items in tests. A significant number of students were deprived of achieving more than 80% scores due to the presence of flawed items in the tests. Similarly, 28 students were placed in low achievement group who deserved to be placed in moderate achievement group. Although awarding grades is not the prime objective of assessment, differentiating high and low achieving students is always. The presence of flawed items blurred the actual boundaries of achievement groups, high performing students appeared as moderately performing and moderate performing students appeared as low performing, compromising the authenticity of such decisions in assessment.

The reliability of a test is an estimate of proportionate amount of random error in the data[14]. In this study there was decrease in the reliability of 9 standard test (tests 1,2,3,5,6,7,8,9,10) after removal of flawed items from the tests. There was increase in the reliability of one (test-4) after removal of flawed items. Most of these differences were small and statistical not significant in this study. The length of an examination and performance of items on test are two important determinants of the reliability of a written test[14]. The psychometrics of flawed items were comparable to standard norms of acceptability and removal of these items decreased the length of the different tests and consequently reduced the reliability of standard test.

There was high rate of flawed items in the tests ranging from 7% to 52% (Mean 22%). There were six most common item writing flaws; negative stem (25%), implausible distractors (19%), unfocussed stem (12.5%), unequal length of distractors (7.8%), none of above (6.6%) and all of above (6.6%) accounting for 78% of all flaws (Table-1). These results closely resembling the finding of Steven M Downing who found five most common flaws accounting for 90% of all flaws ("unfocused stem" or a "negative stem", "all of the above" or "none of the above" options and "partial K-type" items) in his study[8]. Similar finding were also found by Marie Tarrant and James Ware in their study where they found eight types of most common flaws accounting for 85 % of all violations. These included 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[9].

The significant presence of these flaws shows lack of faculty training in item writing. The faculty in our

medical colleges, though highly trained in their respective disciplines, has little or no training in educational assessment methods. The opportunities for such training are few and there is limited number of trained medical educationist in the country. There is also no requirement by regulatory authorities for such training. Writing a quality MCQ is not only difficult and time consuming, but also needs awareness of item writing principles and supervised training. It is only by experience and training that faculty will be able to develop high quality MCQs. These common item writing flaws can easily be rectified by creating opportunities and imparting some focused training in item writing during faculty development workshops in medical institutions. The presence of these item writing flaws is an important construct irrelevant threat to the validity of test results. The training of faculty for item writing and pre-examination item review for correction of these flaws will improve validity of test results[15].

Conclusion

There were unintended negative consequences of using flawed items during assessment. The results of high, moderate and low achieving students were affected by the presence of flawed items in the tests. The findings in the study have demonstrated that more endeavors are needed for quality improvement of in house developed MCQ tests.

Acknowledgements: Prof. Dr. Lubna Baig for her guidance in planning and statistical analysis of the study

Disclaimer: This article is a part of MHPE Dissertation

Conflict of Interest: None to declare

Funding Disclosure: None to declare

References

1. Downing S M. Written Tests. In: Downing S M, Yudkowsky R (editors). Assessment in Health Professions Education. NY. Routledge Taylor and Francis; 2009.p 149-184.
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; 38:974–9.
4. DiBattista D, Kurzawa L. Examination of the Quality of Multiple-choice Items on Classroom Tests. Canad J Scholarship Teach Learn. 2011;2.
5. Case S M and Swanson D B. Constructing Written Test Questions For the Basic and Clinical Sciences. 3rdEd (revised). Philadelphia: NBME; 2003. . (Online) 2014 http://www.nbme.org/publications/item-writing-manual.html
6. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines. Appl Measurement Educ. 2002; 15 (3):309–33.
7. Jozefowicz R F, Koeppen B M, Case M, Galbraith R, Swanson D and Glew R H. The Quality of In-house Medical School Examinations. Acad Med. 2002; 77: 156–161.
8. Downing SM. Construct-irrelevant variance and flawed test questions: do multiple-choice item writing principles make any difference? Acad Med. 2002; 77 (10):103–4.
9. Jozefowicz R F, Koeppen B M, Case M, Galbraith R, Swanson D and Glew R H. The Quality of In-house Medical School Examinations. Acad Med. 2002; 77: 156–161.
10. Downing SM. Construct-irrelevant variance and flawed test questions: do multiple-choice item writing principles make any difference? Acad Med. 2002; 77 (10):103–4.
11. Downing SM. Construct-irrelevant variance and flawed test questions: do multiple-choice item writing principles make any difference? Acad Med. 2002; 77 (10):103–4.
12. 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. 2005; 10: 133–43.
13. Wadi M, Rahim A F A, Yusoff M S B, Baharuddin K A. The effect of MCQ vetting on students' examination performance. EduMed Jo. 2014; 6 (2): 16-26.
14. Axelson RD, Kreiter C D. Reliability. In: Downing S M, Yudkowsky R (editors). Assessment in Health Professions Education. NY. Routledge Taylor and Francis; 2009.p 57-73.
15. Downing S M, Haladyna T M. Validity and its threats. In: Downing S M, Yudkowsky R (editors). Assessment in Health Professions Education. NY. Routledge Taylor and Francis; 2009.p 21-51.

HISTORY	
Date Received	11-03-2019
Date sent for Reviewer	22-03-2019
Date Received Reviewer's Comments	29-05-2019
Date Received Revised Manuscript	25-06-2019
Date Accepted	28-08-2019

AUTHOR'S CONTRIBUTION:

- **Abdul Khalid Awan**
  Conception, Study Designing, Planning, Study Conduction, Interpretation, Analysis, Discussion, Manuscript Writing and Critical Review