

Original Article

EFFECTIVENESS OF A FACULTY DEVELOPMENT PROGRAM ON THE QUALITY OF A-TYPE MULTIPLE CHOICE QUESTIONS IN UNDERGRADUATE MEDICAL AND DENTAL EDUCATION: A QUASI-EXPERIMENTAL STUDY

Amina Gul¹, Nowshad Asim², Muhammad Saleh Faisal³, Muhammad Ihtesham Khan¹,
Muhammad Salman⁴, Sumaira Javed⁵

¹Department of Pathology, Khyber Medical College, Peshawar

²Institute of Health Professions Education & Research, Khyber Medical University, Peshawar

³Department of Pharmacology, Khyber Medical College, Peshawar

⁴Department of Orthopedics, Lady Reading Hospital, Peshawar

⁵Department of Anatomy, Ayub Medical College, Abbottabad

ABSTRACT

Objectives: This study aimed to evaluate the impact of structured FDP on the quality of A-type MCQs constructed by health educators for undergraduate examination.

Materials and Methods: A single-group pretest-posttest quasi-experimental study was carried out in a public sector health professions institution. A one-day FDP was organized for 15 faculty members using NBME (National Board of Medical Examiners) item writing guidelines along with the Revised Bloom's Taxonomy. Complete paired data were obtained from the participants, each submitted five MCQs, both prior to and post the intervention. The MCQs were evaluated using a 15-criterion rubric with score ranging 15-75. The primary outcome measure was the difference in the mean total quality score at faculty level. Other outcome measures included changes in the proportion of cognitive levels and decrease in item writing flaws. Paired T test and Cohen's d were applied for comparative and statistical analysis.

Results: Ten of the fifteen faculty members submitted complete sets of pre and post intervention MCQs, resulting in a total item count of 100 for analysis. There was an increase in the mean total quality score from 55.44 pre-intervention to 56.72 post-intervention, equating to a mean gain of 1.28 with a p-value of 0.003. There was also a corresponding increase in the percentage of higher-order MCQs from 34% to 50%, while flawed items decreased from 74% to 56%. Significant improvements were seen only in curriculum alignment ($p=0.014$) and single-concept focus ($p=0.018$) among other criteria. Significant gains were also recorded in both structural quality ($p=0.040$) and content validity ($p=0.042$) after the FDP.

Conclusion: Participation in a structured FDP resulted in notable improvements in MCQ quality, increased cognitive level of the questions, and decreased item writing flaws.

Key words: Assessment Quality, Dental Education, Faculty Development Program, Medical Education

Cite as: Gul A, Asim N, Faisal MS, Khan MI, Salman M, Javed S. Effectiveness of a Faculty Development Program on the Quality of A-Type Multiple Choice Questions in Undergraduate Medical and Dental Education: A Quasi-Experimental Study. Journal of Khyber College of Dentistry Sep 2026, Vol. 16, No. 3. <http://doi.org/10.33279/jkcd.v16i03.1042>

Correspondence:

Muhammad Saleh Faisal

Associate Professor

Department of Pharmacology, Khyber Medical College,
Peshawar

Email: drsalehfaisal@gmail.com

Date Submitted: May 2026

Date Revised: August 2026

Date Accepted: September 2026

INTRODUCTION

Assessment plays an essential role in medical and dental education, influencing teaching methods, learning outcomes, institutional policies, and accreditation¹. One of the most common assessment methods employed in education comprises mul-

multiple-choice questions (MCQs). They are favored due to their objectivity, reliability, feasibility, and economy². Nevertheless, the quality of the results generated by MCQs largely depends on the quality of question preparation. Despite the prevalence of multiple-choice tests in medical education, there is substantial evidence indicating that many of these questions are flawed. Poor quality MCQs may include ambiguous stems, distractors that do not make sense, cues indicating the right answer, grammatical errors, and low cognitive level³. Such item-writing flaws have adverse consequences for psychometric characteristics, such as validity, reliability, difficulty index and discrimination ability⁴.

MCQs not only assess students' academic performance but also ensure adequate competency of future physicians⁵. Therefore, the National Board of Medical Examiners (NBME) has issued guidelines for writing items effectively and there is a cognitive hierarchy as per Bloom's revised taxonomy^{6,7}. However, there is a paucity of such stringent assessments in low- and middle-income countries, including Pakistan. Several regional studies have reported issues with the structure of undergraduate assessments despite the availability of item-writing guidelines in some institutions⁸. Inadequate institutional support, insufficient standardized training programs, lack of medical education department's expertise and feedback lead to the gap between the recommended practices for assessment and the quality of test items. Consequently, medical and dental colleges struggle to deal with sub-standard MCQs development, reusing poor questions, and cognitive validity differences^{9,10}.

The accreditation agencies including the Pakistan Medical and Dental Council and World Federation for Medical Education increasingly emphasize the need for robust assessment governance and faculty development as an integral aspect of quality assurance¹¹⁻¹³. Also, there is an increasing need for colleges to ensure that their faculties are competent in writing MCQs, else they will continue to develop inappropriate questions¹⁴⁻¹⁶. The role of faculty development programs (FDPs) is highly significant for updating faculty about new developments in item writing practices, ensuring relevance to course learning objectives, and the construction of higher-order cognitive items¹⁷.

While several empirical studies have confirmed positive results following FDPs, most have considered structural analysis or psychometric measures. Very few have utilized paired quasi-experimental designs coupled with multiple criterion rubrics derived from NBME guidelines. The current study focuses on assessing the impact of an organized faculty development session on the improvement of A-type MCQs quality developed by medical and dental faculty. This includes investigating the changes in cognitive level representation as well as reduction in item construction errors using checklist based expert review according to guidelines provided by the NBME and Bloom's revised taxonomy.

MATERIALS AND METHODS

A single group, quasi experimental study was carried out at the Department of Pathology, Khyber Medical and Dental College, Peshawar to assess changes in MCQ quality following a structured faculty development program. Ethical approval for the study was granted by the Institutional Ethics Review Committee vide letter No. 656/DME/KMC, 2025. A purposive sampling strategy was used to recruit 15 faculty members involved in developing A-type MCQs for third-year MBBS examinations. Eligible participants were those actively engaged in undergraduate assessments while individuals holding advanced qualification in Health Professions Education such as Diploma or Masters were excluded. Faculty members who submitted complete sets of pre and post intervention MCQs (five MCQs at each time point) were retained for analysis. Participants with incomplete paired submissions were excluded from the analysis to maintain the integrity of within-subject comparisons.

Faculty Development Intervention

The FDP consisted of a one-day interactive workshop focused on principles of high-quality MCQ construction. Course content incorporated item writing standards of NBME and the revised taxonomy by Bloom. The sessions centered on issues such as typical problems faced during item writing, designing effective and focused stems, creating believable and uniform distractors, removing any grammatical or technical hints, aligning items with curriculum objectives, and involving cognitive processes. Workshop sessions comprised of practice-based tasks, small group work, and feedback.

Data Collection Procedure

Prior to the faculty development session, the participants were required to submit five A-type multiple-choice questions corresponding to learning objectives of Year-2 BDS and Year-3 MBBS curriculum (<https://kmu.edu.pk/examination/guideline/3>). Following the session, five MCQs constructed by each participant were submitted to the researchers (post-FDP phase). Two expert raters, with expertise in health professions education, independently rated all the MCQs on the basis of a standardized scoring criterion. The raters were unaware of whether items belonged to the pre or post training phase, and faculty-level pairing was maintained for all analyses.

MCQ Quality Rubric

The quality of MCQs was evaluated using a 15-criterion rubric based on the NBME guidelines for test question construction and Blooms' updated taxonomy of cognitive behavior. The rubric consisted of two domains: Structural Quality (10 criteria), included clarity of stem, Quality of lead-in, Positive statement, Homogeneity among options, Plausibility, Absence of grammatical/technical cues, Logical sequence, and Stem length. Content Validity and Cognitive Alignment (5 criteria), included Curriculum alignment, Singularity of idea/concept, Clinical context, Scientific accuracy, and Cognitive depth. The score for each criterion was based on a five-point rating scale (1 – poor; 5 – excellent), where the Total Quality Score ranged from 15 to 75. Bloom's Taxonomy Level, Higher Order Categorization, and presence of IWF were recorded separately as dichotomous variables. To facilitate agreement between raters, both reviewers conducted a pilot evaluation of five sample items prior to independent item review. Blinding was maintained throughout to minimize reviewer bias.

Primary outcome measure was the mean change in quality score of test for each faculty. Secondary measures were the changes in percentage of high order questions ($\geq$ Bloom 3 level), presence of IWFs and difference between scores at domain and criterion levels.

Statistical Analysis

The analyses involved aggregating the data by faculty level ($n = 10$). Total pre and post FDP quality scores were computed from 5 items per

respondent. The assumptions of normality for differences between paired means were tested with the Shapiro-Wilk test. With all normality assumptions satisfied, paired samples t tests were utilized to examine the changes in the total quality scores. Statistical significance was considered at $p < 0.05$. The effect size was calculated using Cohen's d statistic (0.2 = small, 0.5 = medium, and 0.8 = large). Higher order MCQs and IWF proportions were subjected to paired comparisons but were not included in the calculation of total quality score to preserve construct validity.

RESULT

Fifteen faculty members participated in the study of which ten provided full data for pre-and post-intervention MCQs. However, five members did not submit complete paired MCQ sets and were therefore excluded from the analysis. The final dataset comprised 100 MCQs, with each participating faculty member contributing five items before and five after the FDP. All MCQ scores ranged between the expected rubric limits of 15 to 75. Binary measures, such as higher-order classification and Item Writing Flaws, were separately recorded and were not included in the total quality score.

Total MCQ Quality Score

At the faculty level ($n = 10$), there was a significant increase in the total quality score post-FDP. This involved an increase in the mean score from 55.44 pre-intervention to 56.72 post-intervention, equating to a mean gain of 1.28 as shown in Figure 1. This change was associated with a 95% confidence interval of 0.58 to 1.98, and a paired t-test showed a p-value of 0.003. Due to the small variation in difference scores, the effect size was calculated as Cohen's $d = 1.30$.

Cognitive Representation and Item Writing Flaws

There was an increase in the number of MCQs aimed at assessing higher order cognitive representation from 34% prior to the FDP to 50% after the FDP, an improvement of 16%. There was a decrease in the number of items having one or more item writing flaws from 74% before the FDP to 56% after the FDP, an improvement of 18%.

Criterion Level Assessment

Out of the total 15 MCQ evaluation criteria, only

two were found to have a statistically significant change in their average scores as depicted in Table 1. The average score for criterion related to curriculum alignment was increased by an average of 0.32 points with p value of 0.014. For the single concept focus criterion, it had a statistically significant increase in average score of 0.30 points with p value of 0.018.

Domain-Level Assessment

The changes in both of the domains assessed are presented in Figure 2. There were significant gains made in both structural quality and content validity after the FDP. The average structural quality score improved from 3.718 prior to the FDP to 3.794 after the FDP, a mean gain of 0.076 that was statistically significant ($p = 0.040$). Content validity score also gained significantly ($p = 0.042$), improving from 3.652 to 3.756 with a mean gain of 0.104.

DISCUSSION

Faculty of undergraduate medical and dental institutes are usually accomplished teachers and scholars; nevertheless, available literature suggest that training can improve their abilities in designing appropriate assessments¹⁸. Faculty along with other duties, must ensure that the assessment tools measure the desired learning objectives⁵. Therefore, faculty development programs FDPs, guidelines, and quality control processes become necessary for achieving valid results.

Results of current study showed that involvement in a systematic FDP contributed to positive changes in the quality of MCQs. While the improvement in the total quality score was modest i.e. 55.44 pre-intervention vs 56.72 post-intervention, the differences were statistically significant and consistent

Table 1: Criterion-Level Improvement Across 15 Rubric Criteria

Criterion	Pre Mean	Post Mean	Mean Difference	SD Diff	t	P Value
Stem Clarity	3.64	3.78	0.14	1.07	0.93	0.359
Lead-in Quality	3.74	3.90	0.16	0.84	1.34	0.185
Positive Construction	3.72	3.86	0.14	1.05	0.94	0.350
Option Homogeneity	3.76	3.70	-0.06	0.98	-0.43	0.666
Plausible Distractors	3.00	3.70	0.70	0.88	0.00	1.000
Absence of Grammatical Cues	3.68	3.84	0.16	1.04	1.09	0.281
Avoidance of Absolute Terms	3.70	3.84	0.14	0.70	1.41	0.164
Absence of Technical Flaws	3.82	3.62	-0.20	0.88	-1.61	0.115
Logical Ordering	3.72	3.86	0.14	1.11	0.89	0.375
Appropriate Stem Length	3.70	3.84	0.14	1.14	0.87	0.391
Curriculum Alignment	3.52	3.84	0.32	0.89	2.54	0.014*
Single Concept Focus	3.64	3.94	0.30	0.86	2.46	0.018*
Clinical Context Integration	3.82	3.84	0.02	0.98	0.14	0.886
Scientific Accuracy	3.66	3.70	0.04	0.83	0.34	0.735
Depth of Cognitive Processing	3.62	3.46	-0.16	1.11	-1.02	0.314

*Statistically significant at $p < 0.05$

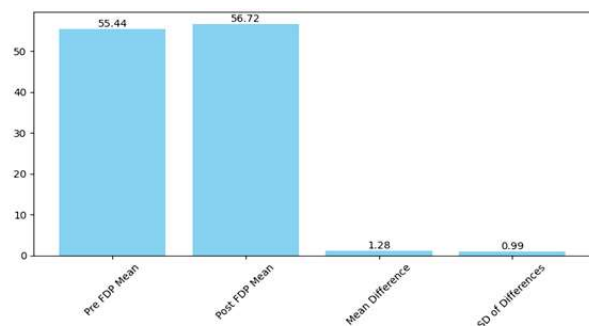


Fig 1: Pre-FDP vs Post-FDP Mean MCQ Quality Scores and Mean Difference

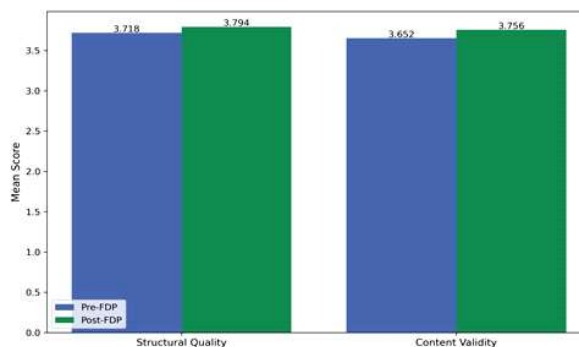


Fig 2: Domain-level Improvements in MCQ Quality (Pre-Post FDP)

across faculty. Higher order thinking skills were also positively influenced i.e. 34% pre-intervention vs 50% post-intervention together with item-writing errors, which fell by 18%. While improvements were more significant in criteria levels for curriculum alignment and single concept focus, enhancements in several microstructural criteria, such as distractor plausibility and decreased reliance on technical cues, were not statistically significant. This is consistent with previous FDP evaluations, highlighting that educators improve more rapidly in conceptual criteria than technical micro-level modifications¹⁹.

Findings of this study have also been backed up by global literature describing the effect of faculty development on item quality. Previous studies, some even done at the regional level, have established the link between targeted training in improving the structural aspect of MCQ items while mitigating typical errors among the faculty members when provided expert guidance^{20,21}. Short-term seminars as well as extended FDP programs have been related to positive results in item writing especially when active engagement is encouraged and personalized feedback is provided²². Notably, Gupta et al. in spite smaller impact reported improvement in distractor homogeneity and total quality score after a short seminar²³. Hye et al. found similar findings, with a substantial increase in MCQ performance following a multiday FDP, indicating that extended training or repeat training might prove more effective than a single training session²⁰.

In the present study, a gain of 16% higher-order cognitive items represent a significant step towards applying and analyzing questions. A local research also confirmed similar trend where a one-day workshop held at the Sindh Institute of Urology and Transplantation has reported the reduction in recall questions along with improved stem quality, leads, and distractors ordering²⁴. Nevertheless, our comparatively modest shift may result from conducting a one-day exercise since higher-order cognitive transformations typically takes time for reinforcement.

The decreased in item writing flaws in our study is consistent with results from other studies, although the baseline flaw rate observed in our institution was relatively higher. In previous studies, the frequency of pre-intervention flaws in MCQs ranged from 50 to 66%, compared to the 74% observed in current

study. Despite a notable post-FDP reduction to 56%, the proportion of flaws remained higher than rates reported in international researches^{25,26}. These variations might be explained by the differences in organizational environment, training quality, and processes of test item creation. Moreover, variability in test scoring methods and definitions of item writing flaws makes direct comparison difficult.

Despite encouraging results, there are some limitations that should be taken into account. First of all, the research was conducted using a single-group design without any control group; therefore, the cause-and-effect relationship is difficult to establish. Secondly, due to small number of participants, the findings cannot be generalized. Moreover, Long-term retention of faculty learning following the FDP was not assessed. Attrition in terms of number of participants from 15 to 10 leads to the possibility of selection bias because only the most interested in completing the whole evaluation process could participate. It should also be noted that even though the use of rubrics to assess quality of assessments provides valuable information, no psychometric measures (such as item difficulty and discrimination) were considered, as well as no impact on students' learning or reliability of the test was evaluated.

CONCLUSION

The results support the role of structured faculty development program as initial step toward quality improvement and cognitive alignment of A-Type Multiple Choice Questions. To maintain and build on the observed gains, it is necessary to employ subsequent measures, including peer review boards, moderated item pools, regular refresher courses, and mentoring programs for novice faculty members in test construction. Further studies should adopt a multi-center design including control groups, and analyze the relationship between rubric-based improvement and test psychometrics as well as student learning outcomes.

REFERENCES

1. Lee HY, Yune SJ, Lee SY, Im S, Kam BS. The impact of repeated item development training on the prediction of medical faculty members' item difficulty index. *BMC medical education*. 2024;24(1):599.
2. Eleragi AM, Miskeen E, Hussein K, Rezigalla AA, Adam MI, Al-Faifi JA, et al. Evaluating the multiple-choice questions quality at the College of Medicine,

- University of Bisha, Saudi Arabia: a three-year experience. *BMC medical education*. 2025;25(1):233.
3. Yakout SM, Hadi K, Alolayan M, Alasmari MM, Alhaji KM, Alobaid LM, et al. Analysis Of Difficulty Index, Discrimination Index And Distractor Efficiency In Multiple Choice Questions. *Journal of Advanced Zoology*. 2024;45(3).
4. 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. *Advances in Health Sciences Education*. 2005;10(2):133-43.
5. Khan HF, Qayyum S, Beenish H, Khan RA, Iltaf S, Faysal LR. Determining the alignment of assessment items with curriculum goals through document analysis by addressing identified item flaws. *BMC medical education*. 2025;25(1):200.
6. Krathwohl DR. A revision of Bloom's taxonomy: An overview. *Theory into practice*. 2002;41(4):212-18.
7. NBME. Constructing written test questions for the basic and clinical sciences (6th ed.). (2021).
8. Awan SL, Manzoor S, Gillani I, Kayani ZA, Farooq H, Khurshid S. Types of item writing flaws in multiple choice questions in medical education and their effect on passing rate of students. *Pakistan Journal of Physiology*. 2023;19(4):49-52.
9. Karthikeyan S, O'Connor E, Hu W. Barriers and facilitators to writing quality items for medical school assessments—a scoping review. *BMC medical education*. 2019;19(1):123.
10. Singh G, Singh R, Ananthakrishnan N. Analyzing the quality of multiple-choice questions in microbiology for second-year MBBS students. *MGM Journal of Medical Sciences*. 2024;11(4):651-58.
11. Haider A, Omar N, Ghafoor K, Mahmood F, Ali M, Muddassar S. Faculty development: Need assessment survey in a newly established Medical College of Lahore. *The Professional Medical Journal*. 2024;31(08):1235-41.
12. Moineau G, León-Bórquez R. WFME recognition programme. *Medical teacher*. 2025;47(1):168.
13. Wajid G, Sethi A, Khan RA, Aamir HS. World federation for medical education: appropriateness of basic medical education standards in Pakistan. *Pakistan journal of medical sciences*. 2019;35(5):1185.
14. Rudolph MJ, Daugherty KK, Ray ME, Shuford VP, Lebovitz L, DiVall MV. Best practices related to examination item construction and post-hoc review. *American journal of pharmaceutical education*. 2019;83(7):7204.
15. Ambusaidi M. Test Item Writing Competence Among Oman College of Health Sciences Nurse Faculty: Widener University; 2022.
16. Qureshi S, Khan R. Challenges faced by faculty of medical education due to the structural variation in its departments across medical colleges. *Pak J Med Health Sci*. 2023;17(01):148-150.
17. Guraya SY, Chen S. The impact and effectiveness of faculty development program in fostering the faculty's knowledge, skills, and professional competence: A systematic review and meta-analysis. *Saudi journal of biological sciences*. 2019;26(4):688-97.
18. Bligh J, Brice J. Further insights into the roles of the medical educator: the importance of scholarly management. *Academic Medicine*. 2009;84(8):1161-65.
19. Ayub R, Kiran F, Shabnam N, Rauf A, Fatima F. Effects of a long term faculty development program on improvement in quality of MCQs: an impact evaluation study. *BMC medical education*. 2025;25(1):541.
20. Hye T, Hossain MF, Roni MA. Interventions to Improve the Quality of Multiple-Choice Questions: A Systematic Review and Meta-Analysis. *American journal of pharmaceutical education*. 2025:101872.
21. Rauf AR, Sultana S. Effect of faculty training on quality of Multiple Choice Questions. *Rawal Medical Journal*. 2021;46(2):430.
22. Nemec EC, Welch B. The impact of a faculty development seminar on the quality of multiple-choice questions. *Currents in Pharmacy Teaching and Learning*. 2016;8(2):160-63.
23. Gupta P, Meena P, Khan AM, Malhotra RK, Singh T. Effect of faculty training on quality of multiple-choice questions. *International Journal of Applied and Basic Medical Research*. 2020;10(3):210-4.
24. Beg MA, Tabassum T, Ali S. Role of faculty development workshop for improving MCQS quality in basic medical sciences. 2021.
25. Nedeau-Cayo R, Laughlin D, Rus L, Hall J. Assessment of item-writing flaws in multiple-choice questions. *Journal for nurses in professional development*. 2013;29(2):52-7.
26. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*. 2006;26(8):662-71.

CONFLICT OF INTEREST
Authors declare no conflict of interest.
GRANT SUPPORT AND FINANCIAL DISCLOSURE
None declared.

AUTHORS' CONTRIBUTION

The following authors have made substantial contributions to the manuscript as under:

Conception or Design: AG, NA, MSF, MIK, MS, SJ

Acquisition, Analysis or Interpretation of Data: AG, NA, MSF, MIK, MS, SJ

Manuscript Writing & Approval: AG, NA, MSF, MIK, MS, SJ

All the authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.



Amina Gul, et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License, which permits unrestricted use, distribution & reproduction in any medium provided that original work is cited properly.