Educ Res Med Sci. 2022 December; 11(2):e119114. https://doi.org/10.5812/erms-119114.

Published online 2023 February 27. **Research Article**

The Role of Qualitative and Quantitative Feedback on Faculties' Quality of Writing Multiple Choice Questions

Amir Shiani [1], Seyed Mojtaba Ahmadi [2], Ghobad Ramezani [3], Fatemeh Darabi [4], Forough Zanganeh [3] and Farhad Salari [5, 3, *]

[1]Department of Speech Therapy, School of Rehabilitation Sciences, Kermanshah University of Medical Sciences, Kermanshah, Iran
[2]Department of Clinical Psychology, School of Medicine, Kermanshah University of Medical Sciences, Kermanshah, Iran
[3]Education Development Center, Kermanshah University of Medical Sciences, Kermanshah, Iran
[4]Department of public health, Asadabad School of Medicine Sciences, Asadabad, Iran
[5]Department of Immunology, School of Medical, Kermanshah University of Medical Sciences, Kermanshah, Iran

[*]*Corresponding author*: Education Development Center, Kermanshah University of Medical Sciences, Kermanshah, Iran. Email: f.salari@kums.ac.ir

Received 2021 August 28; **Revised** 2023 January 10; **Accepted** 2023 January 30.

Abstract

Background: Multiple choice questions (MCQs) are the most common questions in clinical tests. Content validity and appropriate structure of the questions are always outstanding issues for each education system. This study aimed to evaluate the role of providing quantitative and qualitative feedback on the quality of faculty members' MCQs.

Methods: This analytical study was conducted on Kermanshah University of Medical Sciences faculty members using the total MCQs test at least two times from 2018 to 2021. The quantitative data, including the validity of the tests, difficulty, and discrimination indices, were collected using a computer algorithm by experts.

Results: The second analysis revealed that 14 (27.5%) faculty members had credit scores below 0.4, which was within the acceptable range for the overall validity of the test. The results showed a higher difficulty index in the second feedback than the first (0.46 ± 0.21 vs 0.55 ± 0.21, P = 0.30). No significant difference was found in the discrimination index (0.24 ± 0.1.25 vs 0.24 ± 0.10, P = 0.006). Furthermore, there were no significant differences in terms of taxonomy I (61.29 ± 20.84 vs 59.32 ± 22.11, P = 0.54), II (29.71 ± 17.84 vs 32.76 ± 18.82 P = 0.39), and III (8.50 ± 16.60 vs 7.36 ± 14.48, P = 0 .44) before and after feedback.

Conclusions: Based on the results, the questions were not ideal regarding Bloom's taxonomy standards and the difficulty and discrimination indexes. Furthermore, providing feedback alone is not enough, and proper planning by the educational and medical development centers' authorities is required to empower the faculty members in this area.

Keywords: Multiple Choice Questions, Bloom's Taxonomy, Qualitative and Quantitative Analysis

1. Background

Well-assessment is one of the most critical factors in improving the quality of each education system. Multiple choice questions (MCQs) are generally the most common type of questions used in clinical tests. Content validity and the appropriate structure of the questions are always significant issues for test developers. Therefore, it is impossible to distinguish between weak and strong students without observing the structural rules and the appropriate taxonomy level in designing these questions. Low-quality test also reduces learners' motivation, and teachers' and the educational system's efforts will be wasted (1). On the other hand, the type and quality of the test affect the teaching method and the teacher's credibility. Therefore, it is necessary to be careful in preparing questions and performing tests to have the desired characteristics of standard tests, such as validity, reliability, and practicality (2). In this context, educational systems should make appropriate interventions to assess the adequacy of the tests. There is a difference in the quality of four-choice questions in universities regarding structure and learning levels. Various studies have been conducted, including evaluating the quality of multiple-choice tests in a semester of medical school at Mazandaran University of Medical Sciences. Out of 1471 questions related to 25 tests, 64% had one or more structural defects, and most were at the first level of Bloom's taxonomy (3). Baghaei et al. (4) concluded that most questions (84.6%) had one taxonomy level. According to the difficulty index, most questions (332 items) were complex. Among the studied subjects, medical-surgical 3 was the most difficult (61.42%), and obstetric nursing (2) was the least challenging (10%). Regarding the discrimi-

nation index, most questions had an average discrimination coefficient (29.36%), and mental illnesses nursing (1) had the best coefficient among the subjects. Most questions (1.84%) had appropriate structure (4). Shakurnia et al. found that the average difficulty index of the MCQs was 0.59 ± 0.25, and 46.2% had a practical difficulty. The average of the discrimination index of the MCQs was 0.25 ± 0.24, and 57.3% of the MCQs had a discrimination index. Accordingly, combining the two difficulty and discrimination indices showed that only 248 MCQs (30.7%) were ideal. A total of 1525 distractor options (62.9%) were functional distractors (FD), and 889 (37%) were non-functional distractors (NFDs). The results showed that the MCQs should be improved (5). Meanwhile, the analysis of the questions of the specialized midwifery courses of the same university was desirable (3). Shakoornia et al. showed that more than half of the questions designed by the Jundishapur University of Medical Sciences faculty had a correct structure (6). In addition, improving the quality of multiple-choice questions led to an increase in students' level of knowledge (4). Meayari and Biglarkhani indicated that 65.2% of the questions lacked the problems of the overall structure before the intervention. After the intervention, this rate reached 82.8%, which was a significant difference. In 2009, 38% of questions with high taxonomy were designed; in 2010, 53.1%, the differences were also significant. Therefore, intervention can effectively improve the design quality of multiple choice questions as feedback and compliance with technical principles in medical education, even for experienced designers in the design of questions (7). The importance of evaluating students' end-of-term and designing appropriate questions and the lack of knowledge of multichoice questions designed by faculty members is undeniable. Therefore, the need for appropriate interventions in various fields, such as empowerment, continuing education, and feedback, quantitative and qualitative, is felt more than ever at Kermanshah University of Medical Sciences.

2. Objectives

This study aimed to evaluate the role of quantitative and qualitative feedback on end-of-semester tests on faculty members' question design quality in 2018 - 2020.

3. Methods

This analytical study was conducted using a trend impact analysis. The samples were the multiple-choice questions (MCQs) of the medical exams, designed by the faculty members of Kermanshah University of Medical Sciences (KUMS), Iran, in 2018 - 2021, who had delivered final exams and students' results for at least two times via convenience sampling.

The present study included only MCQs test of introductory (non-specialized) science courses. Furthermore, Professors whose questions were analyzed for the first time and had no previous history of feedback of quantitative and qualitative analysis nor training in the field of designing standard tests were selected. The cases examined for the second time were the sample questions of the same professors who had received feedback on quantitative and qualitative analysis.

The quantitative data, including difficulty and discrimination indices, were collected using a computer algorithm. The data interpretation was carried out based on total test validity (0.4 - 1), mean difficulty index of the test (0.0 - 3.7), mean discrimination index (0.2 - 1), measurement criteria error, and a fair number of questions. Experts collected the qualitative data in the fields based on the percentage of the questions with taxonomy I (45%), taxonomy II (40%), and taxonomy II (15%).

3.1. Statistical Analysis

The data were analyzed in SPSS software version 18 for Windows (IBM Corp., Armonk, N.Y., USA), using descriptive statistics (mean, distribution frequency tables, standard deviation) and inferential statistics (dependent *t*-test or Wilcoxon test).

4. Results

In total, 51 faculty members who submitted their MCQs test at least two times to the Educational Development Center (EDC) for the qualitative and quantitative analysis were included in this study. The demographic characteristics of the study participants are shown in Table 1.

The analysis of the questions showed that the credit score of 14 (27.5%) faculty members were below 0.4, and the acceptable range for the overall test validity is 0.4 - 1. Among those with acceptable credit scores, 18 cases (35.3%) were in the range of 0.4 - 59, and 11 (21.6%) and 8 cases (15.6%) were in the range of 0.6 - 0.79 and 0.8 - 1, respectively.

Regarding the mean difficulty index, one individual (2%) achieved a difficulty index in the range of 0.85 - 1 (simple), while 10 (19.6%), 19 (37.3%), 9 (17.6%), and 12 subjects (23.5%) achieved a difficulty index of 0.65 - 0.849 (moderate), 0.449 - 0.649 (difficult), 0.25 - 0.449 (very difficult), and 0 - 0.25 (extremely difficult), respectively. As for the mean discrimination index (acceptable range: 0.2 - 1), 36 individuals (70.6%) achieved a discrimination index of below 0.2, whereas 15 subjects (29.4%) had a discrimination index

Table 1. Frequency and Percentage Distribution of Age, Gender, and Work of Faculty Members Surveyed (n = 51)

Variables	Frequency (%)
Age	
35 - 45	19 (37.3)
45 - 55	25 (49)
55 - 65	7 (13.7)
Sex	
Male	38 (74.5)
Female	13 (25.5)
School distribution	
Medical	23 (45.1)
Dentistry	1 (2)
Pharmacology	8 (15.7)
Health	3 (5.9)
Nursing	7 (13.7)
Paramedical	9 (17.6)
Total	51 (100)

within the acceptable range. Regarding measurement criteria error (acceptable range: 4 - 7), 50 faculty members (98%) were in the range of 0 - 4. Data from 51 individuals who submitted their questions to the EDC for the first and second time were analyzed to investigate the effect of providing feedback to faculty members. The mean taxonomy I, II, and III was estimated at 61.17 ± 22.02, 27.62 ± 16.85, and 9.84 ± 18.03, respectively. The difficulty index in the second feedback analysis was higher than in the first (0.46 ± 0.21 vs 0.55 ± 0.21, P = 0.30). No significant difference was found in the discrimination index (0.24 ± 0.1.25 vs 0.24 ± 0.10, P = 0.006). Furthermore, there were no significant differences in terms of taxonomy I (61.29 ± 20.84 vs 59.32 ± 22.11, P = 0.54), II (29.71 ± 17.84 vs 32.76 ± 18.82 P = 0.39), and III (8.50 ± 16.60 vs 7.36 ± 14.48, P = 0 .44) before and after feedback.

The results showed a significant difference between the two groups regarding the mean Difficulty Index variable, demonstrating that the difficulty rate reduced after providing feedback (Table 2).

5. Discussion

In the present study, the majority (about 80% of the subjects) of the faculty members used taxonomy I question much more frequently than 45% as a standard in MCQ tests, and the rates of taxonomy II and III were 27.62 and 9.84, respectively, which were lower than standard MCQ tests. Faculty members failed to adhere to the standard domain for assessing students' knowledge, and a combination of questions was selected instead. As a result, the students were primarily evaluated regarding their memorization skills. Therefore, evaluations should be designed with adherence to a proportional level of the taxonomy. Only 45% of exam questions could be based on memorization skills, and the remaining items should contain practical and conceptual aspects.

Observations from the literature review suggest this is a common issue. Haghshenas et al. showed that most exam questions were designed based on taxonomy I, which could assess students' knowledge and memory (2). In another study, Pourmirza Kalhori et al. stated that most exam questions had taxonomy I level (8).

The study examined most questions with taxonomy I, so education development centers should empower teachers to design questions with a higher taxonomy. In general, questions designed with taxonomy I, II, and II mostly measure the respondents' knowledge and memory, comprehensive understanding of the lesson, and comprehension of the applicability of the course, respectively (9).

Derakhshan et al. showed that the structural forms of each question were 0.58 ± 0.02 in the pretest and 0.44 ± 0.02 in post-test. Hence, there was a significant difference before and after training using an independent *t*-test. Therefore, holding empowerment programs for faculty members can effectively reduce the number of structural defects of questions and shows the need to maintain and expand such programs in medical education (10). Owolabi et al. (2021) revealed that the program was implemented to strengthen faculty members' quality in designing multiple-choice questions (11). Abdulghani et al. showed that faculty members' longitudinal development workshops help improve teachers' MCQ question writing skills, which also leads to high levels of student competence(12).

According to the results, the exam difficulty index was within the range of 1.3 - 34.6%. Since the optimal difficulty index is in the range of 30 - 80% (8), it could be stated that most of the exams held in most KUMS courses had a suitable difficulty. In Pourmirza Kalhori et al., most evaluated questions had a moderate difficulty index (8). On the other hand, Vafamehr and Dadgostarnia, reported that most questions had a low difficulty index (13). Ashraf Pour et al. claimed that only 38% of the questions had a suitable difficulty index (14). According to Hosseini Teshnizi et al., most questions had an appropriate difficulty index, while the other questions had a low difficulty index and were easy to answer (15).

Moreover, Mitra et al., reported that 80% of the questions had a low difficulty index and were easy (16). Mishmast Nehy and Javadimehr, evaluated the exam questions of the second semester of 2010 - 2011, reporting that 45.6%

Table 2. Comparison of MCQs Quantitative and Qualitative Characteristics of the MCQ Exams Before and After Feedback to Faculty Members (n = 51)

Variables	Before Feedback		After Feedback		Z [b]/T [c]	P-Value
Validity of the tests	0.51 ± 0.23	0.48 (0.36)	0.55 ± 0.22	0.56 (0.27)	-1.04	0.30 [c]
Difficulty Index	0.46 ± 0.21	0.51 (0.35)	0.55 ± 0.21	0.58 (0.36)	-2.88	0.006 [c]
Discrimination Index	0.24 ± 0.1	0.25 (0.11)	0.24 ± 0.10	0.25 (0.13)	-0.19	0.84 [c]
Taxonomy I (%)	61.29 ± 20.84	65 (20)	59.32 ± 22.11	65 (31)	0.60	0.54 [c]
Taxonomy II (%)	29.71 ± 17.84	30 (25)	32.76 ± 18.82	35 (26.25)	-0.84	0.39 [b]
Taxonomy III (%)	8.5 ± 16.60	0.00 (10)	7.36 ± 14.48	0.00 (10)	-0.77	0.44 [b]

Abbreviation: IQR, inter quartile range.
[a] Values are expressed as mean ± SD and median (IQR).
[b] Wilkakson test.
[c] Paired *t*-test.

of the questions had a low difficulty index, whereas 40% had a suitable difficulty index (17). In Sim and Rasiah, 75% of exam questions had a low difficulty index (18). Different types of assessment questions might cause inconsistencies between the mentioned findings. Nevertheless, a similarity between most of these studies was the need to train and upgrade faculty members to design questions with appropriate difficulty, especially in medical and paramedical fields.

In the present study, the discrimination index of the questions was 22.7 - 73.3%. A higher discrimination index indicates the question's discernment, and the closer the index gets to 100, the higher its suitability becomes (16). Notably, the discrimination index of the test items in the study was moderate in most courses and could distinguish weak students from strong ones. In addition, students at different levels of education maintained the ability to respond to more vulnerable students. In Pourmirza Kalhori et al., most exam questions' discrimination index was moderate (8). Hosseini Teshnizi et al. also reported that 57.7% of exam questions had a proper discrimination index (15). According to Ashraf Pour et al., the mean discrimination index of exam questions was 0.14, which showed the improper discernment ability of the questions (14). In Mitra et al., 67% of exam questions had a discrimination index higher than 0.2, considered acceptable (16). According to Shaban and Ramezani, the discrimination power of 40.6% of exam questions was lower than 0.2 even after an educational intervention (19).

In the study by Sanagoo et al., the discrimination coefficient of all 12 tests was low, and the highest discrimination power was 0.32 (1). The discrepancies between the mentioned findings could be due to the different nature of the studies. Similar to the difficulty index, the discrimination index requires the education and improvement of faculty members to design questions with a higher discrimination index. The credit score for the exams of 29% of the faculty members was lower than 0.4. Meanwhile, the acceptable range of a test's credibility is 0.4 - 1. In other cases, 58% of those with acceptable credit scores had a credit of 0.4 - 0.59, while 26% and 15% had credit scores of 0.6 - 0.79 and 0.8 - 1, respectively. Therefore, most university teachers use objective questions to measure students' academic achievement, while there should be more conceptual and practical questions in academic exams. Students' memorization skills may be evaluated only by objective questions, which are less conceptual. Students' conceptual and functional needs should be considered more based on their academic level.

Regarding the measurement criteria error, the acceptable range was 4 - 7 in the present study, while most of the faculty members achieved scores of 0 - 4. Data analysis indicated a difference in difficulty index between the groups when providing feedback to faculty members on 51 people who submitted their first and second questions to the center. After giving feedback, the difficulty of the questions decreased.

5.1. Conclusions

Medical and paramedical fields require accurate assessment due to their high sensitivity. Graduates of these disciplines will be directly involved in maintaining community health after graduation. According to the results, the evaluated exam questions were not ideal regarding Bloom's taxonomy standards and the difficulty and discrimination indexes. More conceptual and practical questions should be designed, and more attention should be paid to the taxonomy levels based on the criteria. Feedback to teachers should also emphasize this topic. In addition, providing feedback alone isn't enough, and the educational and medical development centers' authorities need to plan appropriately to empower faculty members in this area.

Acknowledgments

We extend our gratitude to the Medical Education Studies and Educational Development Center (EDC) of Kermanshah University of Medical Sciences and the university faculty members for assisting us in this research project.

Footnotes

Authors' Contribution: Study concept and design, Farhad Salari; Analysis and interpretation of data, seyed Mojtaba Ahmadi; Drafting of the manuscript, Amir Ahiani & Ghobad Ramezani; Critical revision of the manuscript for important intellectual content, Ghobad Ramezani, Fatemeh Darabi, and Forogh Zanghaneh.

Conflict of Interests: We declare that all authors are faculty members or employees of EDC. There is no conflict of interest to declare.

Ethical Approval: The work was approved by the Ethics Committee of the Kermanshah University of Medical Sciences with code IR.KUMS.REC.1399.1101 (ethics.research.ac.ir/ProposalCertificateEn.php?id=176143).

Funding/Support: This study was supported by the Kermanshah University of Medical Sciences, Iran.

References

1. Sanagoo A, Jouybari L, Ghanbari Gorji M. [Quantitative and Qualitative Analysis of Academic Achievement Tests in Golestan University of Medical Sciences]. *Res Med Educ*. 2010;**2**(2):24–32. Persian.
2. Haghshenas M, Vahidshahi K, Mahmudi M, Shahbaznejad L, Parvinnejad N, Emadi A. [Evaluation of Multiple Choice Questions in the School of Medicine Mazandaran University of Medical Sciences the First Semester of 2007]. *Strides Dev Med Educ*. 2009;**5**(2):120–7. Persian.
3. Hoseini H, Shakour M, Rezaei Dehaghani A. [Validity, Reliability and Difficulty of Multiple-Choice Questions in Specialized Courses of Final Exams in Bachelor Nursing at Isfahan University of Medical Sciences]. *Educ Dev Judishapur*. 2018;**9**(2):129–36. Persian.
4. Baghaei R, Naderi J, Shams S, Feizi A, Rasouli D. [Evaluation of The Nursing Students Final Exam Multiple-Choice Questions in Urmia University of Medical Sciences]. *J Nurs Midwifery*. 2016;**14**(4):291–9. Persian.
5. Shakurnia A, ghafourian boroujerdnia M, Khodadadi A, Ghadiri A, Amari A. [Analytical Study of Quantitative Indices of Multiple-choice Questions of Immunology Department in Ahvaz Jundishapur University of Medical Sciences]. *Educ Dev Judishapur*. 2018;**9**(2):72–83. Persian.
6. Shakoornia A, Khosravi A, Shariati A, Zarei A. [Survey on multiple choice questions of Jondishathe por Medical University of Ahwaz faculty members]. *The 8th National Congress of Medical Education Kerman*. Kerman University of Medical Sciences. 2007. 44 p. Persian.
7. Meyari A, Beiglarkhani M. [Improvement of Design of Multiple Choice Questions in Annual Residency Exams by Giving Feedback]. *Stride Dev Med Educ*. 2013;**10**(1):109–18. Persian.
8. Pourmirza Kalhori R, Rezaie M, Shojee Moghadam AR, Sepahi V, Memar Eftekhari L. [Correlation of quality and quantity index of multiple choice questions exams of residency promotion in Kermanshah University of medical sciences, 2013]. *J of Clin Res Paramed Sci*. 2015;**4**(1):71–8. Persian.
9. Bloom BS. Reflections on the development and use of the taxonomy. In: Rehage KJ, Anderson LW, Sosniak LA, Bloom BS, editors. *Bloom's taxonomy: A forty-year retrospective*. Chicago, USA: National Society for the Study of Education; 1994.
10. Derakhshan F, Allami A, Ahmadi S. Effect of Faculty Training Programs on Improving Quality of Residency Exams in 2013-2014. *Res Med Educ*. 2015;**7**(1):19–26. https://doi.org/10.18869/acadpub.rme.7.1.19.
11. Owolabi LF, Adamu B, Taura MG, Isa AI, Jibo AM, Abdul-Razek R, et al. Impact of a longitudinal faculty development program on the quality of multiple-choice question item writing in medical education. *Ann Afr Med*. 2021;**20**(1):46–51. [PubMed ID: 33727512]. [PubMed Central ID: PMC8102895]. https://doi.org/10.4103/aam.aam_14_20.
12. Abdulghani HM, Irshad M, Haque S, Ahmad T, Sattar K, Khalil MS. Effectiveness of longitudinal faculty development programs on MCQs items writing skills: A follow-up study. *PLoS One*. 2017;**12**(10). e0185895. [PubMed ID: 29016659]. [PubMed Central ID: PMC5634605]. https://doi.org/10.1371/journal.pone.0185895.
13. Vafamehr V, Dadgostarnia M. Reviewing the results of qualitative and quantitative analysis of MCQs in Introduction to clinical medicine course. *Iran J Med Educ*. 2011;**10**(5):1146–52. Persian.
14. Ashraf Pour M, Beheshti Z, Molook Zadeh S. Quality of final examination in students of Babol Medical University, 1999-2000. *J Babol Univ Med Sci*. 2003;**5**(5):42–7.
15. Hosseini Teshnizi S, Zare S, Solati M. [Quality analysis of multiple choice questions(MCQs) examinations of noncontinuous undergraduate medical records]. *Hormozgan Med J*. 2010;**14**(3):170–7. Persian.
16. Mitra NK, Nagaraja HS, Ponnudurai G, Judson JP. The Levels Of Difficulty And Discrimination Indices In Type A Multiple Choice Questions Of Pre-clinical Semester 1 Multidisciplinary Summative Tests. *International e-Journal of Science, Medicine & Education*. 2009;**3**(1):2–7. https://doi.org/10.56026/imu.3.1.2.
17. Mishmast Nehy GA, Javadimehr M. Analysis of multiple choice questions of the second semester examinations held in Zahedan college of Nursing and Midwifery in 2010-2011. *Life Sci J*. 2013;**10**(3):1045–51.
18. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Ann Acad Med Singap*. 2006;**35**(2):67–71. [PubMed ID: 16565756].
19. Shaban M, Ramezani FB. [Effect of Test Item Analysis of Summative Exams on Quality of Test Designing]. *Journal of Hayat*. 2007;**13**(1):5–15. Persian.