



Multiple Choice Question Metamorphosis—Enhancing Question Quality through Faculty Training: A Pragmatic Study

Ankeeta Menona Jacob^{1*}, Kishan Prasad Hosapatna Laxminarayana²

Received: 07 October 2025; Accepted: 05 August 2026

Visual Abstract

ABSTRACT

Background: Faculty in medical colleges often lack formal training in developing multiple-choice questions (MCQs) that target higher-order thinking skills (HOTS). Poorly constructed MCQs can negatively impact the assessment's effectiveness and the examinee's outcomes, ultimately impacting students' learning. The study aimed to compare item-writing flaws and item-analysis metrics after a faculty-training program among medical college faculty in Karnataka.

Materials and methods: A quasi-experimental study was conducted from March to June 2024 at a private medical college in Karnataka, involving 32 medical faculty members trained to identify item flaws and analyze item-analysis metrics. After the second training program, MCQs prepared and administered by the faculty were analyzed to determine the effect of the training program. Descriptive variables were expressed as percentages and proportions, and inferential statistics were used to determine changes in item flaws and analysis metrics before and after the training programs. Statistical analysis was conducted using MS Excel 365 and SPSS version 29.0, with a significance level of $p < 0.05$.

Results: MCQ questions between two training programs reduced recall questions from 737 (87.1%) to 75 (63.6%). The proportions of analytical MCQs increased from 2.8% (24) to 13.6% (16), application MCQs from 9.1% (77) to 17.8% (21), and evaluation MCQs from 0.9% (8) to 5.1% (6). A significant reduction in the difficulty index from 86.0 (66.67–95.1) to 61.1 (44.4–93.5), an increase in the discriminator index from 23.3 (11.0–34.5) to 34.4 (18.1–46.1), and an increase in zero nonfunctional distractors from 26 (12.2%) to 45 (39.5%) were noted after the second training program.

Conclusion: The training programs significantly improved assessment quality by improving item-analysis scores for higher-order thinking skills in MCQs prepared by the faculty. However, they were not effective in reducing MCQ item-writing flaws.

Keywords: Faculty training, Higher-order thinking, Item analysis, Medical education, Multiple-choice questions.

Journal of The Association of Physicians of India (2026); 10.59556/japi.74.1739

INTRODUCTION

Multiple-choice questions (MCQs) are widely used in medical assessments because of their efficiency, reliability, and standardization.¹ The quality of MCQs affects students' competency levels. Well-constructed MCQs assess higher-order cognitive skills using Bloom's taxonomy and the knowledge domain of Miller's pyramid.¹ Poorly constructed MCQs can negatively affect assessment effectiveness and examinee outcomes, ultimately affecting student learning.² Framing MCQs is challenging, but a carefully curated question bank aids academic institutions in conducting assessments.¹

The formulation of MCQs and items is entrusted to faculty members, but many need to become familiar with standard MCQ guidelines for using higher-order thinking skills. One reason cited for poor MCQ/item formulation is a lack of faculty training, which contributes to flaws in MCQ writing.² Item analysis provides information on the validity and reliability of MCQs and is highly dependent on students' level of preparation.¹ Item-analysis metrics are often

used as indices for optimizing questions and preparing question banks.³ Optimization of MCQs produces more valid results, and pre-examination review based on feedback from previous item-analysis scores can augment the reliability of MCQs in the long term.⁴ These findings highlight the need to review and enhance the quality of test items, particularly by adjusting difficulty levels, improving item discrimination, and increasing distractor effectiveness to foster deeper engagement and learning. This may be possible by developing standard operating guidelines for setting MCQs.⁵

Research in the Indian context, in which item analysis was used to improve faculty skills after a single-day workshop on framing multiple-choice questions, found an improvement in item-writing flaws from 27.7 to 15% and reduced heterogeneity from 27.1 to 5.8%, with little improvement in MCQ writing.¹ Other studies conducted in Iran showed that faculty training in one-best MCQ writing significantly improved MCQ quality, decreasing flaw scores from pretest

to posttest, with common flaws found in the stem and lead-in.⁶ However, very few studies have looked at the residual effect of training programs on higher-order thinking skills related to multiple-choice questions in the Indian context.

Towards this end, a study was undertaken to compare item flaws before and 1 month after a training program for single-best-response, higher-order-thinking-skills MCQs prepared by faculty in a medical college, as well as item-analysis scores—facility value (FV), difficulty index (DFI), discrimination index (DI), nonfunctional distractors (NFDs), and distractor effectiveness (DE)—before and 1 month after the training program.

MATERIALS AND METHODS

Study Design and Participants

The study was an interventional, quasi-experimental study involving 32 consenting medical college faculty members at a private medical college in Karnataka, India. Participants were trained to frame single-best-response, higher-order-thinking-skills multiple-choice questions. The study was conducted for 4 months (1 March 2024–30 June 2024).

Data Collection

MCQ Writing Training Needs Assessment of the Faculty

Item-writing flaws and item-analysis metrics were obtained from the College Learning Management System (LMS). The study considered 889 multiple-choice questions available in the participating departments' question banks in the LMS; the questions were carefully evaluated by one of the authors to ascertain item-writing flaws.

¹Professor, Department of Community Medicine; ²Professor, Department of Pathology, KS Hegde Medical Academy (KSHEMA), NITTE (Deemed to be University), Mangaluru, Karnataka, India; *Corresponding Author

How to cite this article: Jacob AM, Hosapatna Laxminarayana KP. Multiple Choice Question Metamorphosis—Enhancing Question Quality through Faculty Training: A Pragmatic Study. *J Assoc Physicians India* 2026;74(10):55–58.

Because the multiple-choice questions had already been administered to medical students as part of their formative assessments, the built-in item analysis of the administered questions—difficulty index (P), discrimination index (D), and distractor efficiency (DE)—was used to establish baseline item-analysis metrics before the training among the faculty in the participating departments.

Training Intervention

The training was held in two consecutive sessions, 1 month apart, with the same faculty members trained in the first session. The training sessions broadly focused on:

- The principles of writing effective MCQs.
- The development of higher-order thinking skills (HOTS) questions.
- Methods to enhance item-analysis metrics.

The training program was administered in two interactive sessions, each lasting 3 hours. The first session trained the faculty in reducing item-writing flaws, and the second trained them in conducting item analysis. Hands-on training in developing higher-order thinking skills (HOTS) MCQs, such as application, analysis, evaluation, interpretation, critical thinking, case-vignette analysis, scenario-based reasoning, and problem-solving, rather than recall-based questions considered lower-order thinking skills (LOTS), was conducted during the first session. The second session provided hands-on training in conducting item analysis using the LMS of the department's course webpage and Microsoft Excel item-analysis sheets available through the Carlo Excels YouTube channel. The supplied resource link is retained in the source manuscript.

The faculty were provided with handouts at the end of both training sessions on strategies for creating effective MCQs, developing higher-order thinking skills (HOTS) questions, and improving item-analysis metrics. The faculty members were allowed to use any method of their choice to develop MCQs and administer them to students before the next training session. Each trained faculty member was required to complete 5–10 higher-order-thinking-skills questions in their specialty, administer them to medical students, and provide the LMS link or anonymized physical copies of the students' answer scripts to the authors for analysis of item-writing flaws and item-analysis metrics after the second training session.

After both training sessions, MCQs created by 27 of the 32 faculty members ($n = 154$) were evaluated for item-writing flaws and item-analysis metrics, including difficulty index (P), discrimination index (D), and distractor efficiency (DE), by one of the authors using standard methods for item analysis.

Statistical Analysis

Descriptive Statistics

The study analyzed faculty MCQ-writing flaws, including approximately 11 major item-writing flaws, rather than subject content per se. These included vague terms, negative statements, and long, complex options in the stem, lead-in, and option set, as shown in Table 1. The results are expressed as percentages and proportions.

The item analysis metrics included pre-training and after the second training program values of the DIF I/difficulty index, discriminatory index/ d value, and distractor efficiency index,

expressed as mean and standard deviation or median and interquartile range (IQR), as appropriate. Non-functioning distractors were graded as poor, moderate, good, or excellent, with percentages and proportions.

Inferential Statistics

For item-writing flaws: The Z test for proportions was used to compare frequencies between the pretraining values obtained from the LMS and the post-second-training-program item-writing flaws among faculty members.

For item analysis: The Mann-Whitney U test was used to compare item-analysis metrics between MCQs evaluated before training and those evaluated after the second training program. For faculty knowledge scores, paired t-test and Wilcoxon signed-rank test were used to compare pretest and posttest scores. The item-analysis metrics of faculty value/difficulty index, discrimination index, and distractor effectiveness in framing MCQs before training and after the second training program were expressed as frequencies and percentages and analyzed using chi-square tests and Z tests for proportions; $p < 0.05$ was considered statistically significant.

For faculty feedback at the end of each training program: Feedback on a 5-point Likert scale collected using Google Forms was expressed as frequencies and proportions. All data were entered in Microsoft Excel and analyzed using SPSS software, version 29.0. A p -value of < 0.05 was considered statistically significant.

Ethics Considerations

The study involved faculty members attending training programs on framing

Table 1: Types of common flaws seen in framing multiple-choice questions before and after the training program among the faculty

Types of flaws in writing multiple-choice questions	Before training ($N = 889$) n (%)	After two training programs ($N = 153$) n (%)	Z value	p-value
Flaws related to the question stem				
Use of vague terms like frequent, most common, never, and always	0 (0)	0 (0)	—	—
The case scenario included is not related to the question	7 (0.8)	4 (2.6)	−2.04	0.041*
Flaws related to lead-in				
Negative statement (which is NOT/FALSE in the characteristic type of question)	24 (2.7)	9 (5.9)	−2.08	0.038*
Questions ending with ALL EXCEPT	137 (15.4)	9 (5.9)	3.14	0.002*
Flaws related to the option set				
Dissimilar length of the answer probe	103 (11.6)	19 (12.4)	−0.30	0.767
Long, complex options	29 (3.3)	4 (2.6)	0.42	0.673
Numbers in the option set are not in ascending/descending order	43 (4.8)	5 (3.3)	0.86	0.393
Use of overlapping numbers in the options ($n = 85$)	17 (1.9)	0 (0)	1.73	0.085
Words could have been added in the lead-in	145 (16.3)	21 (13.7)	0.81	0.420
Unfamiliar acronyms	55 (6.2)	5 (3.3)	1.43	0.152
Use of "none of the above" or "all of the above"	38 (4.3)	4 (2.6)	0.96	0.335

Z test for proportions; * $p < 0.05$ considered statistically significant

single-best-response multiple-choice questions and analyzing student scores. It adhered to ethical considerations and complied with the World Medical Association's Code of Ethics (Declaration of Helsinki) for human experiments. Faculty participants provided informed consent and a participant information sheet outlining the study purpose, risks, benefits, and the right to withdraw without penalty. The study ensured the confidentiality and anonymity of faculty and student data, preventing individual identification in reports or publications. The Institutional Ethics Committee (IEC) approved the study via letter INST.EC/EC/106/2024 dated 5 April 2024.

RESULTS

The training workshop was attended by 32 faculty members from the first to the fourth professional years, ranging from lecturers to professors and department

heads. There were 12 (37.5%) male and 20 (62.5%) female faculty members.

Item-writing Flaws

The faculty members' item flaws before and after the second training program are described in Table 1. Although the frequency of several item-writing flaws decreased after the second training program, a statistically significant reduction was observed in the use of questions ending with "ALL EXCEPT" (15.4% before training vs. 5.9% after training; $p = 0.002$). Though the number of negatively worded and instances where the case scenario was not related to the question apparently increased it may be attributable to the fewer questions evaluated in the post training session. The relative frequency of questions also varied significantly, with improved question quality resulting from a shift from recall (lower-order thinking skills) to application, analysis, evaluation, and case-based scenarios (higher-order thinking skills), as shown in Table 2.

Item-analysis Metrics

The difficulty index decreased from 86.0 (66.67–95.07) before training to 61.11 (44.44–93.38) after the second training program. The discrimination index increased from 23.33 (11.34–34.40) to 34.15 (18.18–45.45). Distractor effectiveness also increased from 36.36 (17.95–52.20) to 43.62 (26.70–62.25) after the second training program. There was a significant increase in the proportion of MCQs with zero nonfunctional distractors, from 26 (12.2%) to 44 (38.9%), and a significant reduction in those with two nonfunctional distractors, from 57 (26.8%) to 19 (16.8%). The proportions of MCQs with three nonfunctional distractors decreased from 83 (39.0%) to 33 (29.2%), while those with one nonfunctional distractor decreased from 47 (22.1%) to 17 (15.0%); these reductions were not statistically significant. Overall changes in item-analysis metrics are presented in Table 3. Table 4 shows a statistically significant increase in posttest scores for item-analysis knowledge

Table 2: Cognitive levels of multiple-choice questions before and after the training program among the faculty

Cognitive level	Pretraining (N = 889) n (%)	After two training programs (N = 153) n (%)	Z test value	p-value
Lower-order thinking (recall-based)	772 (86.8)	98 (64.1)	7.01	<0.001*
Higher-order thinking skills MCQ				
Application	84 (9.4)	26 (17.0)	-2.81	0.005*
Analytical	25 (2.8)	21 (13.7)	-6.07	<0.001*
Evaluation	8 (0.9)	8 (5.2)	-4.02	<0.001*
Case-vignette/scenario-based	68 (7.6)	44 (28.8)	-7.79	<0.001*

Z test for proportions; * $p < 0.05$

Table 3: Item analysis of single-best-response MCQs created by the faculty after two training programs ($n = 27$)

Item-analysis metrics	Pre-training median (IQR) (n = 213)	After two training programs, median (IQR) (n = 113)	Z† value	p-value
Difficulty index	86.0 (66.67–95.07)	61.11 (44.44–93.38)	3.37	<0.001*
Discrimination index	23.33 (11.34–34.40)	34.15 (18.18–45.45)	-4.10	<0.001*
Distractor effectiveness	36.36 (17.95–52.20)	43.62 (26.70–62.25)	-2.21	0.027*
Distractor efficiency	Pretraining (n = 213)	After two training programs (n = 113)	Z value	p-value
3 nonfunctional distractors (poor)	83 (39.0%)	33 (29.2%)	1.8	0.07
2 nonfunctional distractors (moderate)	57 (26.8%)	19 (16.8%)	2.05	0.039*
1 nonfunctional distractor (good)	47 (22.1%)	17 (15.0%)	1.55	0.121
Zero nonfunctional distractors (excellent)	26 (12.2%)	44 (38.9%)	5.7	<0.001*

IQR, interquartile range; †Mann-Whitney U test; Z test for proportions; * $p < 0.05$

Table 4: Knowledge scores of the faculty on item analysis at the end of two training programs ($n = 27$)

Characteristics	Pretest (out of 10)	Posttest (out of 10)	T value	p-value
Mean	4.1 ± 2.2	7.7 ± 2.1	5.83 [#]	<0.001
Range	1–8	3–10		
Median	3 (2–6.5)	8 (7–9)	-3.83 [£]	<0.001

[#]Paired t-test; [£]Wilcoxon signed-rank test; p -value < 0.05

Table 5: Feedback on the training program by the participants ($n = 22$)

Question	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
The workshop content on item analysis was clear and easy to understand	18 (81.8) [§]	2 (9.1)	1 (4.5)	1 (4.5)	0 (0.0)
The workshop provided a good balance between theory and practical application	15 (68.2) [§]	6 (27.3)	1 (4.5)	0 (0.0)	0 (0.0)
The resource persons were knowledgeable and engaging	19 (86.4) [§]	3 (13.6)	0 (0.0)	0 (0.0)	0 (0.0)
After attending this workshop, I am confident that I can conduct an item analysis of MCQs	16 (72.7) [§]	5 (22.7)	0 (0.0)	1 (4.5)	0 (0.0)
I can understand the basics of item analysis in quality-assurance multiple-choice questions (MCQs)	15 (68.2) [§]	6 (27.3)	1 (4.5)	0 (0.0)	0 (0.0)
I can critically analyze MCQs for medical undergraduates that I framed myself using item analysis	15 (68.2) [§]	5 (22.7)	2 (9.1)	0 (0.0)	0 (0.0)
I am willing to conduct an item analysis for MCQs and am prepared to improve the quality of the multiple-choice questions I frame	15 (68.2) [§]	5 (22.7)	1 (4.5)	1 (4.5)	0 (0.0)

[§]represents the highest percentage

among faculty members participating in the workshop.

Feedback from the Faculty ($N = 22$)

Table 5 shows that the training program was successful, with participants finding the content clear, engaging, and practical; feeling confident in applying item-analysis techniques; and being highly motivated to implement their learning.

DISCUSSION

The study shows improvement in item-analysis metrics after training, with marked improvements in difficulty, distractor efficiency, and discrimination indices. Nonfunctional distractors decreased, and the number of questions based on recall decreased. Analytic, application, evaluation-related, and case-scenario-based multiple-choice questions increased. Significant reductions were observed in questions ending with "ALL EXCEPT."

Changes in Item-writing Flaws

The shift from lower-order thinking skills to higher-order thinking skills (HOTS) MCQs was an important outcome of the training. The proportion of recall-based questions decreased from 86.8% (772/889) pretraining to 64.1% (98/153) posttraining, while the proportion of questions assessing application increased from 9.4% (84/889) to 17.0% (26/153), analysis from 2.8% (25/889) to 13.7% (21/153), and evaluation from 0.9% (8/889) to 5.2% (8/153). This shift indicates a move away from predominantly recall-based assessment toward greater representation of higher-order cognitive skills following the training.

Changes in Item-analysis Metrics

The distribution of nonfunctional distractors differed significantly between pretraining

and posttraining items. In particular, there was a significant increase in items with zero nonfunctional distractors, from 26 (12.2%) to 44 (38.9%), and a significant reduction in items with two nonfunctional distractors, from 57 (26.8%) to 19 (16.8%). Items with three nonfunctional distractors decreased from 83 (39.0%) to 33 (29.2%), while those with one nonfunctional distractor decreased from 47 (22.1%) to 17 (15.0%); these individual reductions were not statistically significant.

CONCLUSION AND LIMITATIONS

The training program for faculty members in MCQ construction significantly improved assessment quality by reducing item flaws and increasing higher-order thinking skills. Regular training can improve MCQ assessment through item-flaw reduction and improvement in item-analysis metrics. However, the study's limitations include subjective bias in faculty selection and the need for training sessions to reinforce item writing flaws and item analysis to sustain the training program efforts.

CONFLICT OF INTEREST

None.

SOURCE OF SUPPORT

None.

AUTHOR CONTRIBUTIONS

AMJ: Conceptualization, study design, data curation and analysis, manuscript drafting—original draft, manuscript review and editing, final approval of manuscript; KPHL: Conceptualization, study design, manuscript drafting—original draft, manuscript review and editing, final approval of manuscript.

ARTIFICIAL INTELLIGENCE DISCLOSURE

During the preparation of this work, the authors used Turnitin to estimate the level of purported plagiarized content and QuillBot to paraphrase content flagged as supposedly plagiarized. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the final published article.

ORCID

Ankeeta Menona Jacob  <https://orcid.org/0000-0002-9839-3556>

Kishan Prasad Hosapatna Laxminarayana  <https://orcid.org/0000-0002-4071-5866>

REFERENCES

- Kumar D, Jaipurkar R, Shekhar A, et al. Item analysis of multiple-choice questions: a quality assurance test for an assessment tool. *Med J Armed Forces India* 2021;77(Suppl 1):S85–S89.
- Gupta P, Meena P, Khan AM, et al. Effect of faculty training on quality of multiple-choice questions. *Int J Appl Basic Med Res* 2020;10(3):210–214.
- Bhat SK, Prasad KH. Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: a cross-sectional study. *Indian J Ophthalmol* 2021;69(2):343.
- Chibueze Izah S, Catherine Odubo T, Emeka Ajumobi V, et al. Item analysis of multiple-choice questions (MCQs) from a formative assessment of first-year microbiology major students. *Res Rev Insights [Internet]* 2021;5(2):1–6.
- Chavda P, Misra S, Duttaroy B. Item analysis of multiple-choice questions based on undergraduate assessment in community medicine. 2015;9(1):66.
- Rauf AR, Sultana S. Effect of faculty training on quality of multiple-choice questions. *Rawal Med J* 2021;46(2):430.
- Monrad SU, Bibler Zaidi NL, Grob KL, et al. What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's taxonomy. *Med Teach* 2021;43(5):575–582.
- Mansour T, El Tarhouny SA, Azab M. Assessment of multiple-choice questions test item quality after post-exam item analysis. *Int J Sci Res* 2022;32–34.
- Lee HY, Yune SJ, Lee SY, et al. The impact of repeated item development training on the prediction of medical faculty members' item difficulty index. *BMC Med Educ* 2024;24:599.