J Med Edu. 2020 February; 19(1):e103482. doi: 10.5812/jme.103482.

Published online 2020 June 7. **Research Article**

Impact Measurement on Medical Faculty for Adhering to Appropriate Guidelines in Framing Effective Multiple-Choice Questions for Item Analysis

Surajit Kundu [1, *], Jaideo M Ughade [2], Anil R Sherke [1], Yogita Kanwar [1], Samta Tiwari [1], Ravikant Jatwar [1], Richa Gurudiwan [1] and Sumati G Kundu [3]

[1]Department of Anatomy, Late Shri Lakhiram Agrawal Memorial Govt. Medical College, Raigarh, India
[2]Department of Anatomy, Govt. Medical College, Nizamabad, India
[3]Department of Pharmacology, Late Shri Lakhiram Agrawal Memorial Govt. Medical College Raigarh, India

[*]*Corresponding author*: Department of Anatomy, Late Shri Lakhiram Agrawal Memorial Govt. Medical College, Raigarh, India. Tel: +91-7583828825; +91-8827264640 Email: dr.surajitkundu@rediffmail.com

Received 2020 April 11; **Accepted** 2020 April 22.

Abstract

Background: Multiple-choice questions (MCQs) are the most frequently accepted tool for the evaluation of comprehension, knowledge, and application among medical students. In single best response MCQs (items), a high order of cognition of students can be assessed. It is essential to develop valid and reliable MCQs, as flawed items will interfere with the unbiased assessment. The present paper gives an attempt to discuss the art of framing well-structured items taking kind help from the provided references. This article puts forth a practice for committed medical educators to uplift the skill of forming quality MCQs by enhanced Faculty Development programs (FDPs).
Objectives: The objective of the study is also to test the quality of MCQs by item analysis.
Methods: In this study, 100 MCQs of set I or set II were distributed to 200 MBBS students of Late Shri Lakhiram Agrawal Memorial Govt. Medical College Raigarh (CG) for item analysis for quality MCQs. Set I and Set II were MCQs which were formed by 60 medical faculty before and after FDP, respectively. All MCQs had a single stem with three wrong and one correct answers. The data were entered in Microsoft excel 2016 software to analyze. The difficulty index (Dif I), discrimination index (DI), and distractor efficiency (DE) were the item analysis parameters used to evaluate the impact on adhering to the guidelines for framing MCQs.
Results: The mean calculated difficulty index, discrimination index, and distractor efficiency were 56.54%, 0.26, and 89.93%, respectively. Among 100 items, 14 items were of higher difficulty level (DIF I < 30%), 70 were of moderate category, and 16 items were of easy level (DIF I > 60%). A total of 10 items had very good DI (0.40), 32 had recommended values (0.30 - 0.39), and 25 were acceptable with changes (0.20 - 0.29). Of the 100 MCQs, there were 27 MCQs with DE of 66.66% and 11 MCQs with DE of 33.33%.
Conclusions: In this study, higher cognitive-domain MCQs increased after training, recurrent-type MCQ decreased, and MCQ with item writing flaws reduced, therefore making our results much more statistically significant. We had nine MCQs that satisfied all the criteria of item analysis.

Keywords: MCQs, Item Analysis, Difficulty Index, Discrimination Index, Distractor Efficiency

1. Background

Assessment is an integral part of medical learning and training (1). Assessment tools can be hampered by their poor design, proficiency of users, deliberate abuse, and unintentional misuse. To establish the usefulness of a particular assessment format, the following criteria should be considered, as framed by Van der Vleuten (2, 3): (1) Reliability (a measure of producing consistent results); (2) validity (performance efficiency of a test); (3) affecting future learning and practice; (4) suitability to learners and faculty; and (5) expenses (to the individual trainee, institution, and society at large). For a medical graduate or postgraduate trainee, proper assessment is a dominant motivator, as medical learning involves acquiring skills for future implementation (4). This is because assessment emphasizes the recall of factual knowledge (5) and examines the higher levels of cognitive abilities that have a deep impact on students' choice of the learning approach (6). Medical students can be evaluated and assessed by an array of

methods including Long Essay question (LEQ), Modified Essay question (MEQ), Multiple Choice question (MCQ), Short Essay question (SEQ), Objective Structured Practical examination (OSPE), Objective Structured Clinical examination (OSCE), clinical simulations, and VIVA. Among them, one of the most versatile, valid, and common methods of evaluation is the use of MCQs (1, 6).

Multiple Choice questions (MCQs) were introduced into medical examinations in the 1950s as a reliable method of testing knowledge to replace traditional LEQs (7). Multiple-choice questions can be designed to assess the higher cognitive levels of students in undergraduate and postgraduate medical evaluation (1, 8). Multiple-choice questions can be effectively employed for both formative and summative assessments to screen a wide section of a medical subject and scrutinize a large number of students in lesser time simultaneously with rapid turnaround of results (9, 10). Its acceptance is based on its objectivity, feasibility, high internal consistency, reliability, and accuracy, thus avoiding inter-examiner bias (8). Multiple choice questions can also test all domains of learning (11). Recently, considerable innovative revisions have occurred in undergraduate and postgraduate medical education in the form of development of Objective Structured Practical examination (OSPE) or Objective Structured Clinical examination (OSCE) (3, 12). These revised curricula focus on Millar's conceptual framework of medical competence. These are "knows" "knows how" "shows how" and "does", which are arranged as various layers of Miller's pyramid.

Numerous studies have reported the disappointing quality of MCQs in medical institutes and books of diverse medical subjects available in the market (9, 13, 14). Shah et al. (15) advocated the scrutiny of such MCQs before their use for assessment. Although creating good MCQs is laborious, time-consuming, challenging, and an art to be acquired (1, 10), the evaluation and compilation of results require simple computer assistance (1, 11). Moreover, MCQs have garnered a largely negative reputation. This is based on the belief that a student has to correctly recognize the answer from the list of options, rather than spontaneously generate it (a phenomenon termed "cueing") (16). Furthermore, MCQs are often perceived to be far removed from the real-life demands of the practicing clinician and therefore, less clinically valid. Multiple-choice questions are also susceptible to internal errors and writing flaws, which adversely impact student performance (17-19).

2. Objectives

This study aimed to understand the much-needed art among medical faculty in framing effective MCQs (items) along with the simultaneous need for item analysis, which would serve as a valid tool of assessment to reinforce learning among undergraduate and postgraduate medical students. The data of item analysis in the present study are also to be correlated with values obtained previously by other researchers to decide on the validity of the present manuscript.

3. Review of Literature

As commented by Azer (20), medical faculties often perform tasks for which no formal training is available. Developing and analyzing articles is a responsibility for which they have no experience and training. Framing can result in errors if staff are not sensitive enough to develop test items, and they are not adequately trained, resulting in a lack of the quality of many tests.

As stated by Downing (21), a good test question starts with identifying the most important information or skills for writing the question you need to learn. A direct relationship exists between the educational purpose and the test material. Therefore, the tests must come directly from the objectives; avoid examining the knowledge of the treatment and focus on the relevant content. Controversial elements should be avoided, especially when knowledge is incomplete or information is disputed. It may be easier to determine appropriate test questions by examining the subtopics of an article or other topics and identifying key concepts or concise sentences. From there, the key points can be written as declarative sentences, which create a clear picture of what students need to learn. It has been suggested that the written concept is a clear statement, proposal, or policy as an important part of the instruction as it is worth analyzing.

Steinert et al. (17) in their systematic review on health professions education stated that Faculty Training programs (FDP) are associated positively with teaching effectiveness for both immediate and long-term effects. Faculty training programs for that reason are prime for the formation of valid and reliable assessment materials. Shalini Chandra et al. (9) proposed that knowledge, skill, and the mean score of MCQs quality significantly enhanced in post-training sessions of FDP. Significant improvements in item analysis indices were documented [Df I ($P = 0.001$), DI ($P = 0.02$), and DE ($P < 0.0001$)]. Patil et al. (1) advocated methods to construct quality MCQs. The mean difficulty index,

discrimination index, and distractor efficiency were 38.3%, 0.27, and 82.8%, respectively. Of 30 items, 11 items were of higher difficulty level (DIF I < 30%) while five items were of easy level (DIF I > 60%). A total of 15 items had very good DI. Of the 90 distractors, there were 16 (17.8%) non-functional distractors (NFDs) present in 13 (43.3%) items. Sahoo et al. (8) documented their findings to show that most items failed to be in the acceptable range of difficulty level; however, some items were rejected due to the poor discrimination index. Their analysis helped in the selection of quality MCQs having high discrimination and average difficulty with three functional distractors. They strongly proposed that item analysis procedures should be incorporated into future evaluations to improve the test score and properly discriminate among students. Poornima et al. (13) declared that MCQs are the most widely used tools for the screening of students in competitive exams as part of formative evaluation. The objectivity, the ability to cover a wide range of topics, and the possibility of assessing a large number of students in a short period have made MCQs a versatile tool of evaluation. However, to maintain the quality of the examination system, reliability and validity of MCQs are of utmost importance.

3.1. Hypothetical Questions

How to construct an effective MCQ?

What are the faculty guidelines in framing effective MCQs?

How to evaluate the item and distractor analysis as a valid tool of item analysis and assessment?

3.2. Basic Structure of Multiple Choice Questions

The basic MCQ or "item" is of a single best response type, wherein the examinee attempts to choose only one answer from a set of usually four options provided. An item consists of a "stem", followed by several options. Sometimes, the stem is followed by the "lead-in question". The correct answer in the list of options is called a "key" and the incorrect options are called "distractors". Thus, the basic parts of an ideal single best response type MCQ can be defined as:

1) Stem (vignette): The context around which the question is asked. It can be a short vignette or case scenario.

2) Question (Lead-in): A clearly stated question to indicate what the student has to do.

3) Distractors (Options): The alternative incorrect options to the question.

4) Answer (Key): The correct answer.

This can be illustrated in Figure 1. Clear cut directions for students are a must for effective MCQs. The instructions should be the same for a common set of questions.

MCQs are of different types as classified by Hubbard and Clemans (1971) in the following: (1) One best response type; (2) K-type (combined response MCQs as the next most widely used MCQ type); (3) Matching type; (4) Relationship analysis type; (5) Case history type (patient management problem); (6) Pictorial type; and (7) Multiple independent true-false selection type.

3.3. Suggestive Art of Framing Effective Multiple Choice Questions

Constructing meaningful and worthwhile MCQs is a herculean and time-taking task. Thorndike and Hagen state that "An indigenous and talented item writer can construct multiple-choice items that require not only the recall of knowledge but also the use of skills of comprehension, interpretation application, analysis, or synthesis to arrive at the keyed answer" (22).

researchers have also rightly quoted that "the greater you experience in their construction, longer it takes per item to construct a reasonably fair, accurate, and inclusive question". This means the more you acquire the art of constriction of good quality MCQs, the difficult it becomes to construct them (22).

The art of constructing flawless MCQs can be perfected with repeated practice and optimum patience. Flawed, unfocused, irrelevant MCQs make it easier for students to answer the question correctly, based on their trial and error skills alone instead of cognitive skills and they fail to examine the trait that is the focus of assessment. Flawed MCQs also act as barriers that negatively affect student evaluation. Therefore, the need to develop reliable flawless test items has been documented in the curriculum-based medical education guidelines as proposed by the Medical Council of India, 2019. In the present context of writing MCQs, we included the guidelines by Haladyna and the American National Board of Medical Examiners (NBME) (23, 24).

3.4. Phases in of Framing Multiple Choice Questions

The framing of MCQ is based on three phases: (A) Formation phase; (B) construction phase; and (C) evaluation phase.

3.4.1. Formation Phase

The time at which the general instructions for MCQ exams are defined like what are the topics, what is the purpose of exam, is MCQ the best choice of assessment?

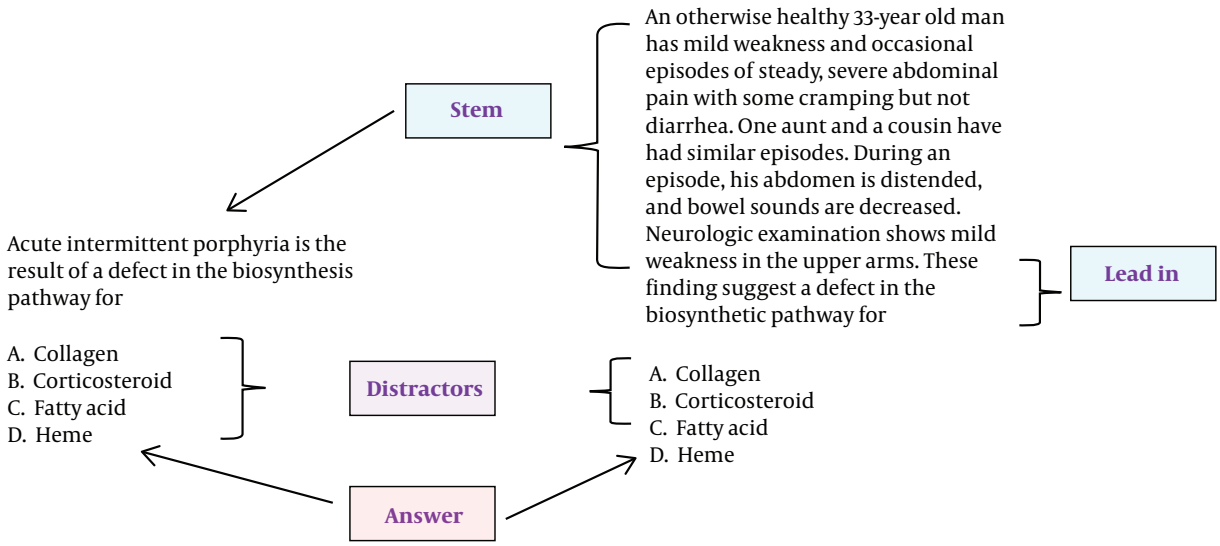

Figure 1. Left question written as combined stem and lead-in; right question written as a case scenario with a separate lead-in (adopted from the American National Board of Medical Examiners Manual, NBME, page 37).

What will be the appropriate number of questions and how much time will be spent on the exam? What can be deducted to qualify for the exam? Does it allow a negative marking? What is the nature of MCQ, for example, is it a type of single best answer or a type of multiple-answer?

The general guidelines to follow during the training phase are as follows: Select important topics to know important things, avoid vague and controversial fields, include broad topics, and choose an appropriate number of questions.

3.4.2. Construction Phase

This phase concentrates on the construction of various parts of an MCQ, with four parts, namely direction, stem, question, distractor and answer (As described earlier).

3.4.3. Evaluation Phase

It is the phase after an MCQ has been constructed and administered to students in an examination. It is done to assess the quality of MCQ. Hence, this phase is also called the posttest phase or the item analysis phase. The purpose of MCQ/item analysis is to answer three simple questions, as follows:

1) Was any MCQ/item too difficult or too easy? (Difficulty index)

2) Did the items discriminate between those students who knew the material from those that did not? (Discrimination index)

3) What is the reliability of the exam?

At the end of exam MCQ are marked and considered for item analysis. The item analysis provides an assessment of MCQ/item's difficulty and several tester's discriminatory skills. The most common indices in this context are (A) Difficulty Index (P); (B) discrimination ratio (d); and (C) Distracter Effectiveness (25).

4. Methods

The present study was planned and conducted as an exclusive part of FDP organized by the Medical Education Unit (MEU) of Late Shri Lakhiram Agrawal Memorial Government Medical College Raigarh (CG).

Informed consent taken from MEU Committee beforehand. To assess the impact of training on developing high-quality single best response MCQs, all faculty members (n = 60) present on the MEU session (for two days) on "framing quality MCQs and item analysis" and 200 MBBS students participating for the Item Analysis of Late Shri Lakhiram Agrawal Memorial Government Medical College Raigarh (CG) were considered as the study sample.

Simultaneously Kirkpatrick's four level of satisfaction comprising of level of satisfaction with the workshop (level 1), level of learning (level 2), behavior change (level 3) and long term impact (level 4) has also been taken into consideration.

4.1. Study Design

This is a questionnaire-based pretest-posttest design observational study. After the study set-up, a faculty satisfaction questionnaire was developed following the guidelines of the Association of Medical Education in Europe (AMME) (9). As the study would be conducted in two days, the departmental meetings were organized beforehand and the faculty willing to participate (with prior written informed consent) were asked to frame 10 MCQs of their specialized medical subject, which were to be discussed on the days of faculty training. Hence, a database of 600 MCQs was framed from 60 faculty members willing to participate and each obtained set of 10 MCQs was earmarked serially. All MCQs were of the single best response type comprising a stem and four choices (one key and three distractors). This database of MCQs was numbered as set I (Pre FDP Set).

To reduce the subjectivity in the evaluation of test items, we used the preformed structured checklists (9, 12) for framing MCQs. They were disbursed among all participating faculty as lecture handouts to test the quality of MCQ items after the intervention. To reduce bias judgment, each faculty was given a separate sheet of 10 MCQs, which were not framed by him/her.

On day 1, the pretest was handed out to the participants to assess their knowledge and perception regarding MCQ framing and item analysis. This was followed by a didactic interactive lecture session on "framing quality MCQs and item analysis" where all participants were enlightened on framing good-quality single best response MCQs. On day 2, the participants were given a sheet containing 10 MCQs, which were previously submitted before the workshop as discussed in methodology, along with structured checklists for framing MCQs for incorporating any modifications. Thus, these MCQs could be considered as fit for including in MBBS examinations.

The following changes made in the stemswere excluded from the analysis: (A) Change from except (straight font) to except(italics/bold/underline); and (B) Change from Not (straight font) to not (italics/bold/underline) Thus, we had a new modified database of 600 MCQs, which was named as Set II (Post FDP Set).

Subsequently, the participants were requested to fill in the faculty perception questionnaire and posttest questions were administered to assess learning. Both pretest and posttest questions were a set of 10 short answer questions to assess learning before and after the session (Table 1).

Among set I and set II, the same numbered 100 MCQs were randomly selected by peer-reviewed Medical Council of India (MCI) approved trained Medical Education Technology (MET) workshop specialists to be finally distributed to 200 MBBS students (100 from set I and 100 from set II). 100 MCQs of the set I formed pretest MCQs and 100 MCQs of the set II formed posttest MCQs. The specialists used structured checklists as designed by previous authors (12) to review the quality of MCQ items before and after FDP to assess the test items. The checklist consisted of 21 markers for assessing MCQ scores. Each marker was allotted one mark making the total marks to 21. Each item was reviewed according to the checklist and the final scores were calculated out of 10 by the unitary method of mathematics. Thus, the calculation of marks according to a structured checklist could be utilized for the mean MCQ test score of pretest and posttest (Table 1).

Each correct response was awarded one mark. No mark was given for blank responses or incorrect answers. There was no negative marking. Thus, the maximum possible score of the overall test was 100 and the minimum was zero. The obtained data were entered in MS Excel 2016 and the analyzed scores of 100 students were categorized into high scoring (H) group (top 33%), mid scoring (M) group (middle 34%), and low scoring (L) group (bottom 33%) after arranging the scores of participating students in descending order. Thus, 100 MBBS students were categorized into the high achieving group (H) and low achieving group (L) as described above (20 each).

To judge the effectiveness of an item (MCQ), a total of 100 MCQs and 300 distractors (100 $\times$ 3) were analyzed and based on the results, various item analysis indices including difficulty index (DIF I; Df I), discrimination index (DI), distracter efficiency (DE), and non-functional distracter (NFD) were calculated for each item, as follows:

$$DIFI = [(H + L)/N] \times 100 \quad (1)$$

$$DI = [(H - L)/N] \times 2 \quad (2)$$

where, N is the total number of students in both groups, and H and L are the numbers of correct responses in the high and low scoring groups, respectively.

The first two item analysis indicators are guiding principles when a university plans to create a large question bank for online examinations. For the calculation of these indices, the following steps are usually taken as standard (13):

A) Arrangement of the entire test scores into ascending or descending order of performance.

B) Dividing the test scores into quartiles, i.e. four parts.

C) Analysis using the upper (high scores or high achievers) and lower (low scores or low achievers) quarters (13).

Table 1. Checklist for Framing Quality Multiple Choice Questions

Areas	Do's	Don't
1. Content related	Each item should focus on important content area/learning outcome	Avoid opinion-based items.
	Ensure each item is independent of the others	Avoid tricky items.
	Frame items that are specific, clearly-defined, and clinically oriented and include the main idea in the question.	
	Avoid nonsense words, unreasonable, unnecessary, and irrelevant statements.	
	Choose items satisfying the level of difficulty index.	
	Questions should correlate with principles, rules, or facts in a real-life context.	
	Frame questions interpreting cause-effect relationships.	
	Avoid answering one question in the test by giving the answer somewhere else in the test.	
	Frame questions relating the student's ability to justify methods and procedures.	
	Give clear instructions.	
2. Writing the stem	Ensure that the directions in the stem are very clear.	Avoid negatives such as NOT or EXCEPT; if used, ensure that the word appears capitalized, boldfaced, and underlined.
	Always include the central idea and common elements in the stem.	Avoid double negatives.
	Keep it clear, concise, and unambiguous.	
	Items with lead-in should indicate how to answer the MCQ.	
	Keep vocabulary simple and in a defined statement to be completed by one option.	
	May be a direct question or an incomplete statement.	
	Students should be able to understand without reading it several times, including the distractors (Put as much question as possible in the stem).	
	Avoid vague expressions like fairly high, considerably greater, etc.	
	Avoid clues suggestive of the right answer	
	Avoid extremes, never, always, only, etc.	
	The stem should not ask for an opinion.	
3. Writing the distractor	Choices should not be overlapping but independent and or mutually exclusive.	Avoid all-of-the-above.
	One and only one of the choices should be the right answer.	None-of-the-above should be used carefully.
	The content of choices must be homogeneous (Functional distractors should be of the same category as the correct response).	Avoid grammatical inconsistencies
	Avoid negatives such as NOT.	
	All distractors must be plausible even if the number of options per question changes.	
	Avoid the use of specific determiners such as always, never, completely, and absolutely.	
	Use answers given in previous open-ended exams to provide almost realistic distractors.	
	Change the location and sequence of the correct answer according to the number of choices.	
	Balance the placement of the correct answer, i.e. arrange distractors in a logical or numerical order, wherever appropriate.	
	The length of choices must be almost similar including numerical observations.	
	Distractors that act as misconceptions and common student errors are very effective.	
	Avoid giving clues.	
4. Formatting and style-related	Format the item vertically instead of horizontally.	Avoid abbreviations (Short forms)
	The entire item should be on the same page.	
	Mix an optimum number of items of different levels of difficulty.	
	Group similar formats together.	
	Have the test reviewed by someone who can find mistakes, clues, grammar, and punctuation problems (Proofreading is a must.).	
	Must be grammatically correct, punctuation, capitalization, and spelling (simple, precise, and unambiguous wording).	
	Keep it brief and minimize the amount of reading in each item.	

The described method simplifies the overall calculations, hence speeding up the analyses. The extreme test scores could differentiate a well-prepared candidate from an unprepared candidate. The minor modification of

MCQs had a direct influence on the extreme scores more than it did on the average scores. The middle two quarters representing the average or median scores are generally not considered in the calculation of difficulty index or discrimination ratios and are influenced by factors like guessing, examination stress, individual personality, logic, risk-taking, and confusion and thus may not test the knowledge of the participants. However, when the number of students taking the examination is less than 30, the entire score set can be divided into two halves, upper half and lower half scores (13).

Item analysis methods and calculations (Working definition and formula)

An MCQ satisfying all the three criteria of item analysis (Dif I, DI, and DE) was considered an ideal/good quality MCQ.

4.2. Statistical Analysis

The data were analyzed in Microsoft excel 2016 software. The Student's t test was used to assess the level of significance. Besides, the statistical significance was specified as $P < 0.05$. In addition, the researcher adopted the Cohen's d test for calculating the effect size to compare the scores obtained at pretests and post-tests to assess the learning of trainees (participants) at pre-training and post-training (9, 12).

5. Results

The faculty perception of FDP was evaluated according to Kirkpatrick's level of evaluation with questionnaires and responses (up to level 3), as depicted in Figure 2.

5.1. Interpretation

5.1.1. Level 1 (Reaction)

Analyzed with filled faculty satisfaction questionnaires. Results showed an average rating of 4 to 5, which pointed to the faculty satisfaction with FDP on a Likert scale of 1 - 5.

A total of 100 MCQs having 300 distractors (incorrect responses) were analyzed among 60 MBBS students. The mean score and standard deviation were 79 .7 and 4.6, respectively. The total score out of 100 ranged from 23 to 89 (23% to 89%). For evaluating, the students were arranged in descending order from the highest score 89 to the lowest score 23. The first 33.33% of the students were in the high achieving group and the last 33.33% of the students were in the low-achieving group. The middle 33.33% of the students were excluded from the calculation.

5.1.2. Level 2 (Learning)

To analyze the pre-test and post-test responses to assess learning, 10 SAQs were administered on both pre and posttests. The level was also analyzed by the MCQ score obtained by the per unitary method as explained in methodology (out of 10).

The mean pretest score was 2.56 whereas the mean posttest score was 9.21 ($P = 0.001$). The effect size was 0.90, which was large based on Cohen's classification (0. 2 = small; 0.5 = medium, and 0.8 = large). This reflects significant learning from FDP (Table 2).

Table 2 also displays the mean scores and effect size of MCQs to put forth a transparent and clear improvement in the quality of MCQs as decided based on a checklist scoring pre-training versus post-training sessions. The quality of MCQs before training was definitely low (5.55 ± 0.556) but it was significantly higher in post-training (8.86 ± 2.56) with the effect size of 0.84, which signals a medium to large effect size.

Level 3 (behavior change or transfer): This level was assessed by comparing the scores of MCQs (as explained in Table 1) and different indices of item analysis pre-training versus post-training. The results are tabulated in Tables 2-4 and Figure 3. The flow chart showing the evaluation based on Kirkpatrick's levels is displayed in Figure 4.

Table 4 reflects the following interpretations:

1) Difficulty Index (Dif I): (A) 24% items and 14% items were in Dif I of < 30 in pre and post training sessions respectively; (B) moderate MCQs with Dif I of 31 - 40 included only 38% in the pretest, which increased to a major percentage of 70% with post-training FDP; (C) easy MCQs with Dif I > 60% were drastically reduced from 38% in pre-training to 16% with post-training FDP.

2) Discrimination Index (DI): (A) Only 10% of MCQs could discriminate between high and low achieving groups in pre-training MCQs but among post-training items, the percentage increased to 25% for good DI MCQs (DI > 0.40); (B) recommended MCQs (with or without changes) with DI in the range of 0.20 - 0.39 increased by a major percentage from 30% to 57% due to FDP training sessions; (C) Thirty eight percent of the MCQs had DI in the discarded MCQ range of < 0.19 in pre-training sessions and 16% in post-training sessions; (D) MCQs with negative DI dramatically decreased from 20% to only a meager value of 2% with FDP training sessions.

3) Distractor Effectiveness (DE): (A) Items with 0 NFD, i.e. DE of 100%, were seen in 18 and 61 items in pre- and post-training MCQs, respectively; (B) sixty-six point sixty-six percent of DE, i.e. items with one NFD, were interpreted

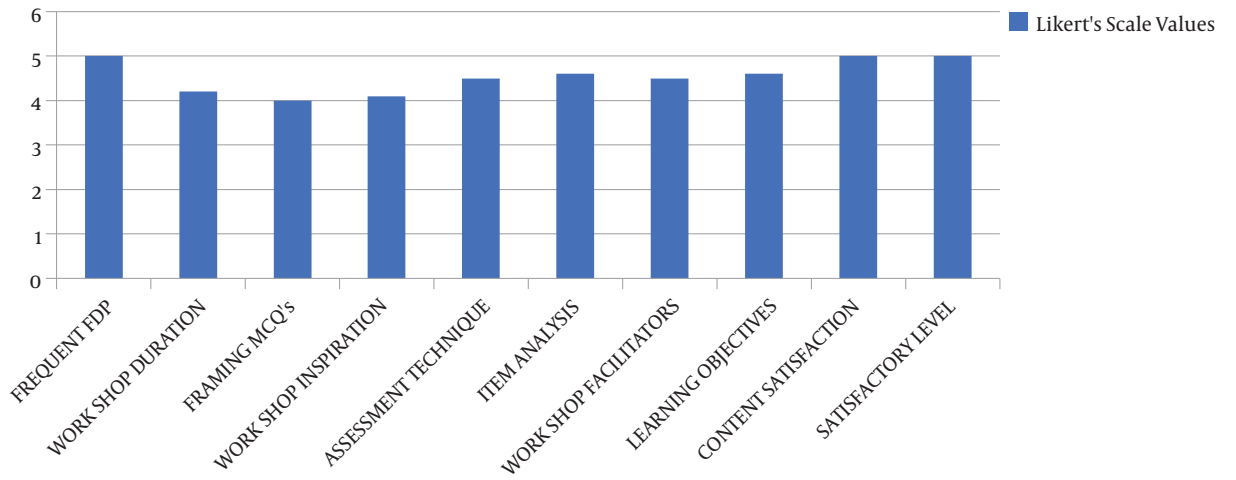

Figure 2. Faculty perception of FDP

Table 2. Participants' Pre and Posttest Scores to Assess Learning (N = 60)[a]

Pretest and Posttest Scores to Assess Learning			
Pretest Mean	**Posttest Mean**	**P Value**	**Effect Size**
2.56 ± 0.548	9.21 ± 0.195	0.001	0.90
Mean MCQ Score as Per Checklist			
Mean MCQ Score (Pretest)	**Mean MCQ Score (Posttest)**	**P Value**	**Effect Size**
5.55 ± 0.556	8.86 ± 2.56	0.001	0.84

[a]Values are expressed as mean ± SD.

in 40% and 27% of the MCQs during pre- and post-training changes, respectively; (C) items with two NFDs, i.e. MCQs with DE of 33.33%, were seen in 35% and 11% of the MCQs in pre and post-training sessions, respectively; (D) Zero DE items, i.e. MCQs with three NFD, reduced from 7% to 1% with FDP.

The deeper dissection of DE in Table 4 shows a clearer picture of the effectiveness of distractor due to the changes incorporated in the items as a consequence of FDP, which was invariably reflected as changes in pre- and post-training MCQs. These modifications are put forth in Table 5.

Considering 100 MCQs with one correct response and three distractors, the total number of distractors should be 300 (100 × 3). Hence, Table 3 gives the following information: (A) NFD greatly reduced from 131 to only 52 in pre and post-training MCQs, respectively; (B) functional distractors increased from 169 to 248 due to FDP sessions among MCQs.

A closer relationship has been chalked out to further bring forth the importance of FDP training by the Medical Education Unit with the presentation of "framing quality MCQs and item analysis" with a correlation between NFD, Dif I, and DI. This has been summarized in Table 6.

As visualized in table 6, the total number of NFD in pre and post training sessions were 131 and 52 respectively. Table 6 further clarifies the intricate correlation of NFD with Dif I and DI.

Interpretations:

A) Difficult MCQs had 58 (44.27%) out of 131 NFDs and 10 out of 52 (19.23%) NFDs in pretest and posttest questions, respectively.

B) Moderate MCQs had 62 (47.32) out of 131 NFDs and 36 (69.24%) out of 52 NFDs in pre-training and post-training, respectively.

C) Easy MCQs had 11 (8.4%) out of 131 NFDs and 6 (11.53%) out of 52 NFDs as a pre and posttest comparison, respectively.

Leaving out MCQs with negative and discardable DI characters and considering good MCQs with DI in the

Table 3. The Characteristic Features and Cutoffs of Item Analysis

Parameter	Difficulty Index (DIF I/DiF) or Facility Value (FV)	Discrimination Index (DI)	Distractor Effectiveness (DE)
Formula	[(H + L)/N] × 100; N is the total number of students in both groups. H and L are the numbers of correct responses in the high and low scoring groups, respectively.	[(H - L)/N] × 2; N is the total number of students in both groups. H and L are the numbers of correct responses in the high and low scoring groups, respectively.	The percentage of students having marked the distracter as the right answer.
Characters and features	Describes the percentage of students who select the correct answer to an item. Higher the value of the difficulty Index, easier the question; so a higher value of DIF I means an easy question. It is calculated as the percentage of students who correctly answered the item. It ranges from 0 to 100%. The recommended optimal range of DiF I is 30% to 70%.	It is the ability of an item to differentiate between the high and low achieving groups. It ranges from 0 to 1. If DI is higher, the item has a greater ability to differentiate between high and low achievers. The recommended DI value is > 0.25; the DI value of 0.15 - 0.25 is acceptable with revision whereas the DI value of < 0.15 is discarded. The range can be -1 to +1. If ideal DI is 1 for an item, it exactly discriminates between students of lower and higher abilities. High-performing students select the correct response for an item more often than do low-performing students. If this is true, the assessment is having a positive DI (between 0.00 and +1.00). This signifies that students with a high total score choose the correct answer for a specific item more often than do students who have a low overall score. However, if low-performing students will get a specific item as correct more often than do the high scorers, then that item will have a negative DI (between -1.00 and 0.00). Here, a good student suspicious of an easy question takes a harder path to solve and end up being less successful while a student of lower ability by guess selects correct responses.	It is the effectiveness of incorrect options (distractors) given in the item to be chosen by a participant. It simply shows whether distractors are functional distractors or nonfunctioning distractors (NFD). NFD is an option other than the correct answer which is selected by less than 5% of total students in high and low groups while the distractors that are selected by 5% or more than 5% of the students are considered functional distractors. Based on the number of NFDs in an item, DE ranges from 0 to 100%. If an item contains three, two, one, or no NFDs, then DE should be 0, 33.3%, 66.6%, or 100% respectively.
Categories and Cutoffs	Very difficult: Difficulty index of less than 30%.	Good discriminator: Discrimination index of more than or equal to 0.2	By analyzing the distractors, it becomes easier to identify errors so that they might be removed, replaced, or revised.
	Poor discriminator: Discrimination index of less than 0.2		
	Acceptable: Difficulty index of 30% to 70%.		
	Inference: 0.40 or above: very good item; 0.30 - 0.39: reasonably good; 0.20 - 0.29: marginal item (i.e. subject to improvement); 0.19 or less: poor item (i.e. to be rejected or improved by revision).; (> 0.30 recommended)		
	Very easy: Difficulty index of above 70%.		
	Inference: < 30: difficult MCQ; 31 - 40: good MCQ (moderate); 41 - 60: very good MCQ (Moderate); > 60: easy MCQ		

range of 0.20 - 0.30 and > 0.30, pretest MCQs showed that only 27 (20.61%) out of 131 could discriminate between high and low achieving groups, but the data increased to 40 (76.92%) out of 52 as posttest questions.

Table 7 clearly shows:

A) Pretest MCQs were easier than posttest MCQs

B) Pretest MCQs had a very low power of discrimination among high and low achieving groups

C) Distractor efficiency was very poor for pretest questions

6. Discussion

The present manuscript was advocated with the prime intention of spreading the light of medical education technology among medical teachers and examiners following the guidelines of the Regional Training Centre Medical Council of India so that the Medical Faculty can be more confident in framing MCQs. The objective was satisfied by our sincere attempt to implement the process of item analysis to decide on the validity of MCQs along with distractors, which finally has the prime motive to prepare competent medical graduates in the future.

An ideally constructed MCQ following the basic rules

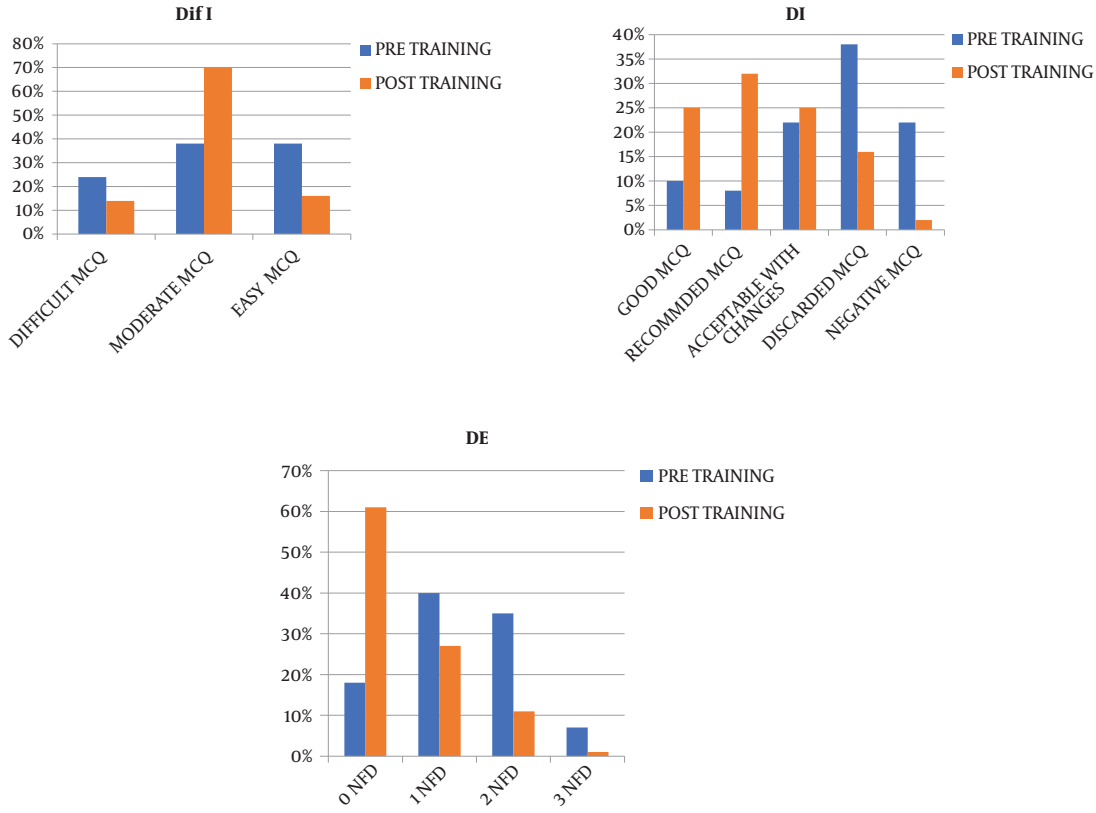

Figure 3. Difficulty index, discrimination index, and distracter effectiveness between pre-training and post-training

of item analysis can be a recommended assessing tool (a means of evaluation) for various types of examination and diverse levels of cognition of Bloom's taxonomy. A well-structured MCQ should have a moderate level of difficulty (> 30% - 60%) with a higher discrimination index (> 0.25) and 100% distractor efficiency, that is, all three incorrect responses should act as distractors. These MCQs are an immediate guiding star for a student of the medical field to sharpen his/her cognitive understanding of the subject and attempt any formative and summative assessment with confidence. Thus, item analysis is an important meticulous instrument for judging the quality of MCQ, as it is beneficial for both the examiner and examinee.

Questions framed by a trained faculty will certainly have an edge over those framed by an untrained faculty in framing MCQs for valid licensing undergraduate or postgraduate medical examinations (so that they are well in the recommended range of difficulty index, discrimination index, distractor efficiency, and presence of nonfunctional distractors. These good MCQs can be the best judge to choose the best competent Indian Medical Graduate (IMG) or Foreign Medical Graduate or Postgraduate students.

Our study initiated with the satisfaction of the participants with the FDP utilizing Kirkpatrick's model of the outcome. We had a very effective workshop as participants rated the FDP with high scores (Figure 2). Simultaneously, 60 participants were immensely satisfied with the literature and handouts about the workshop as a continuous learning method in medical education. A long everlasting preliminary satisfactory impact of participants for any evaluation is a must for any positive changes in the right direction. Many similar studies also reported useful and relevant participant satisfactory levels with FDPs, as found in the present study.

Flawed MCQs (within the stem, key, or distractor) provide clues to the answer, making the MCQ easier, affecting the performance of high achievers and inflating the low achievers, which is always unwarranted. Hence, item analysis not only selects quality MCQs but also removes flawed MCQs. The present study throws ample light on the development of good quality MCQs as reflected on the

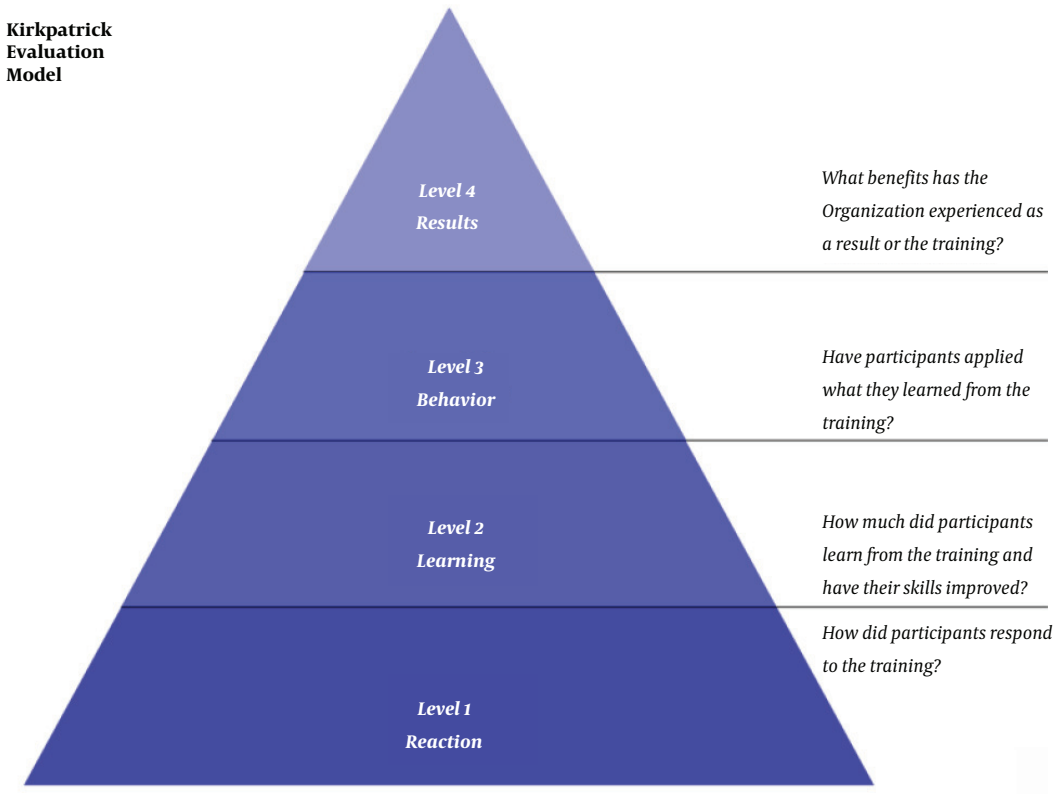

Figure 4. Kirkpatrick's evaluation model (courtesy: lucidchart.com)

changes made by the participants using the guidelines to frame MCQs. Our present research is in congruence with many other studies previously undergone. The changes were classified in results under Dif I, DI, DE, and the number of NFDs by pre and posttest analyses of items.

The methodology attempted in the present manuscript was guided by previous researchers and our results and observations are consistent with many other previous similar studies. The results (as percentages) of the post-training FDP MCQs, which were distributed to the MBBS students, were in the acceptable ranges based on the difficulty index (70%), discrimination index (82%), and distracter effectiveness (61%). Thus, these MCQs and distracters framed by the faculty of Late Shri Lakiram Agrawal Memorial Govt. Medical College Raigarh (CG) can be safely added to our college question bank as they satisfy the criteria of acceptability. Tarrant and Ware (26) and Vik and Ware (27), Chandra et al. (9), Sahoo et al. (8), Patil et al. (1), and Ben-David (28) also showed similar results regarding item analysis.

Ben-David (28) stated that FDP trained faculty had higher mean scores than did untrained doctors in the medical licensing examination questions. The item that poorly discriminates or has high and low Dif I ought to be reviewed by content experts as it would reduce the validity of the test. Our manuscript made an attempt to speak in the same direction.

Again, DI provides us with ready information in context efficiency to differentiate students with high or low proficiency skills and higher the DI is, the more the ability will be to efficiently discriminate. In the present study, 22% of pre-training MCQs had a negative DI, but it strikingly reduced to 2% in the post-training sessions. Items with negative DI are vague, poorly prepared, discardable questions having nonfunctional distractors. Hence, items with negative DIs should be removed from the question bank. The FDP conducted through "framing quality MCCQs and item analysis" in Late Shri Lakhiram Agrawal Memorial Govt. Medical College Raigarh (CG) also gave a positive sign as the recommended percentage of MCQs within the discrimination index increased from 30% to 57% and 82% including the good MCQs. With further intricate searching

Table 4. Difficulty Index, Discrimination Index, and Distracter Effectiveness Between Pre-Training and Post-Training[a]

	Pre-Training Session (N = 100)	Post-Training Session (N = 100)
Difficulty index (Dif I)		
< 30 (difficult MCQ)	24 (24)	14 (14)
31 - 40 (moderate MCQ)	20 (20)	37 (37)
41 - 60 (moderate MCQ)	18 (18)	33 (33)
> 60 (easy MCQ)	38 (38)	16 (16)
Discrimination index (DI)		
> 0.40 (good MCQ)	10 (10)	25 (25)
0.30 - 0.39 (recommended MCQ)	8 (8)	32 (32)
0.20 - 0.29 (MCQ acceptable with Changes)	22 (22)	25 (25)
< 0.19 (discarded or poor MCQ)	38 (38)	16 (16)
Negative	22 (22)	2 (2)
Distractor efficiency (DE)		
100% (items with 0 NFD)	18 (18)	61 (61)
66.66% (items with one NFD)	40 (40)	27 (27)
33.33% (items with two NFDs)	35 (35)	11 (11)
0% (items with three NFDs)	7 (7)	1 (1)

[a]Values are expressed as No. (%).

Table 5. The Number of Functional and Nonfunctional Distracters Among a Total of 300 Distractors (as per Distractor Efficiency in Table 4)[a]

Pretest (Distractor = 300)		Posttest (Distractor = 300)	
Functional Distractor	**Nonfunctional Distractor (NFD)**	**Functional Distractor**	**Nonfunctional Distractor (NFD)**
54	0	183	0
80	40	54	27
35	70	11	22
0	21	0	3
Total = 169 (56.33)	Total = 131 (43.66)	Total = 248 (82.66)	Total = 52 (17.66)
Total distractor = 300		Total distractor = 300	

[a]Values are expressed as No. (%).

of items in Dif I and DI ranges, it was found that the maximum discrimination was visible (n = 49, DI > 0.25) with acceptable levels of difficulty (30% to 70%) during post-training MCQs.

In pre training, only a handful of MCQ's were in the acceptable range of Dif (28). Ware and Vik (27) also approved our observation regarding the DI within 40% to 70% of Dif I.

The results by Halikar et al. (29) on item analysis of 20 MCQs in ophthalmology showed that the percentages of acceptable MCQs based on the difficulty index and discrimination index were 35% and 50%, respectively. All MCQs in their lists had at least one NFD but the percentage of functional distracters was 23%. The authors concluded that item analysis could create a validated bank of MCQs with known values of indices within the recommended range from which question paper setters could choose the appropriate MCQs for an examination.

Namdeo et al. (30) published their paper on item analysis of 25 MCQs in pediatrics and reported that 60% and 68% of MCQs were acceptable based on the difficulty index and discrimination index, respectively. Moreover, 12% of the MCQs had no NFD. 46% of the distracters (incorrect alternatives) of the items they framed were found to be functional. The authors concluded that item analysis is helpful to delineate technical lacunae in MCQs and provide ac-

Table 6. Items with Nonfunctional Distractors and Their Relationships with Difficulty Index (Dif I) and Discrimination Index (DI)[a]

	Difficulty Index (Dif I)	Items (N) with Nonfunctional Distractors (NFD) (N = 100) and Number of NFD = 131	Discrimination Index (DI)	Items (N) with Nonfunctional Distractors (NFD) (N = 100) and Number of NFD = 131
Pretest	< 30 (difficult MCQs)	24 items had 58 (44.27) NFDs	Negative	22 items with 66 (50.38) NFDs
	31 - 40 (moderate MCQs)	20 items had 38 (29) NFDs	< 0.19 (discarded or Poor MCQs)	38 items with 38 (29) NFDs
	41 - 60 (moderate MCQs)	18 items had 24 (18.32) NFDs	0.20 - 0.29 (MCQs acceptable with Changes)	22 items with 16 (12.21) NFDs
	> 60 (easy MCQs)	38 items had 11 (8.4) NFDs	> 0.30 (good MCQs)	22 items with 11 (8.5) NFDs
	Difficulty Index (Dif I)	**Items (N) With Nonfunctional Distractor (NFD) (N = 100) and Number of NFD = 52**	**Discrimination Index (DI)**	**Items (N) With Nonfunctional Distractor (NFD) (N = 100) and Number of NFD = 52**
Posttest	<30 (difficult MCQs)	14 items had 10 (19.23) NFDs	Negative	2 items with 2 NFDs (3.85)
	31 - 40 (moderate MCQs)	37 items had 17 NFDs (32.70%)	< 0.19 (discarded or Poor MCQs)	16 items with 10 (19.23) NFDs
	41 - 60 (moderate MCQs)	33 items had 19 (36.54) NFDs	0.20 - 0.29 (MCQs acceptable with Changes)	25 items with 16 (30.77) NFDs
	> 60 (easy MCQs)	16 items had 6 (11.53) NFDs	> 0.30 (good MCQs)	57 items with 24 (46.15) NFDs

[a]Values are expressed as No. (%).

Table 7. Mean and Standard Deviation (SD) of Difficulty Index, Discrimination Index, and Distractor Efficiency[a]

Parameter	Pretest	Posttest
Difficulty Index	75.66 ± 14.56	56.54 ± 20.55
Discrimination Index	0.18 ± 0.10	0.26 ± 0.12
Distractor Efficiency	38.65 ± 12.45	89.93 ± 19.43

[a]Values are expressed as mean ± SD.

curate information and ways to modify them, appropriately increasing their validity. Item analysis on 50 MCQs in anatomy by Mehta et al. (6) revealed 62% and 70% of MCQs in the acceptable range of difficulty index and discrimination index, respectively. 34% of MCQs had no NFD and 18% of the distracters were functional. The authors concluded that item analysis is a vital tool in the hand of Medical Education Technology (MET) for developing MCQs having higher pedagogic and psychometric values. Our results also focus on a similar direction as previous research. We showed Dif I, DI, and DE were 70%, 57%, 61% (with no NFD), and 27% (with one NFD), respectively.

The examiner often concentrates on choosing a plausible, functional, and appropriate destructor, which is widely accepted as the most difficult part of creating MCQs. Destructor analysis allows us to easily identify the student's response towards NFD. In this survey, with 300 distractors, the percentages of functional and NFD in pre- and post-training were 56.33%, 43.66%, and 82.66% and 17.33%, respectively. Besides, items with zero, one, two, and three NFDs, i.e. DE of 100%, 66.33%, 33.33%, and 0%, were effectively calculated as 18%, 40%, 35%, 7% and 61%, 27%, 11%, 1%, respectively (pre-training and post-training). Finally, the mean Dif I, DI, and DE obtained in our study (56.54, 0.26, and 89.93, respectively (Table 5)) was very congruent with the values of research authored by Gajjar et al. (31), Hingorjo and Jaleel et al. (32), Sim et al. (33), and Vyas and Supe (34).

6.1. Limitations

1) We probably should draft a more rigorous item analysis manuscript following Bloom's taxonomy of cognitive skills.

2) Small sample size.

3) Future Workshops to investigate other easy methods of item analysis.

4) Failure to calculate the long-term impact level of item analysis through FDP.

5) Selection of a large number of MCQs.

6) Focusing more on the evaluation method and failure to calculate internal consistency.

7) Tentative item analysis data influenced by the MBBS students being examined, our instructional procedures, and random errors.

8) The low number of students in high and low achieving groups (20 in each).

9) The present study had only nine MCQs that satisfied all the recommended criteria for item analysis.

10) The need for knowledge, comprehension, application, analysis, synthesis, and evaluation of higher-order MCQs.

6.2. Conclusions

We analyzed the cognitive level and quality of MCQs in writing errors. In this study, higher cognitive-domain MCQs increased after training, recurrent-type MCQs decreased, and MCQs with item writing flaws reduced, making our results much more statistically significant.

Frequent FDP has been proposed in the new curriculum based medical education (CBME).

Despite being a valuable tool, the method of item analysis is not voluntarily adopted and accepted by many Medical Colleges due to the lack of awareness, inappropriate compulsion from regulatory authorities, precious time, and undue labor involved and a pseudo-perception by our Medical teachers that subjective validation of medical students may be sufficient without deeper objective item analysis procedures. It has been well-documented that subjective validation is highly variable from one teacher to another and its sensitivity is relatively low as compared to the standard item analysis procedures. Moreover, nowadays, the use of easy user-friendly downloadable software can significantly reduce the time and labor involved in item analysis. Hence, we should be the torch bearers to shoulder the responsibility of spreading awareness, installing software support, and communicating a clear mandate to the regulators to popularize the procedure of item analysis to increase the validity of medical examination assessment to effectively assess all the three domains of medical teaching.

MCQs can be used as a meaningful and effective assessment tool in medical education. The quality of MCQ depends entirely on the quality of the article and the presence of qualified protesters. Defective MCQs interfere with the evaluation process, and therefore, it is important to develop reliable and valid components that are fault-free at the national level. Preparing multiple choice quizzes requires a lot of time, effort, and commitment to test quality, reliable, high-level thinking skills, and to align with the objectives of the curriculum. To evaluate students' knowledge, we, as medical teachers, need to be proficient in composing effective test materials. We propose the necessity of further research with increases in participating faculty to interpret the long-term impact of the faculty development programs.

6.3. Recommendation

We strongly propose all Medical Schools to implement the simple software-based calculations in item analysis. This not only will delete all the flaws that might have crept in our minds regarding MCQs framing but will also make us more receptive to our part to adopt ourselves and implement the varied highly laudable new methods of medical education and teaching. This will definitely be welcomed with open hearts and minds by our own future medical professionals, which will make them more competent to face the ever-increasing burden of Medicine and make medical learning and evaluation partially stress-free.

Acknowledgments

The authors would like to sincerely thank respected Dean, HOD and all faculty members and staff of the Department of Anatomy, participating faculty in the FDP, Medical Education Unit (MEU), and all participating MBBS students of Late Shri Lakhiram Agrawal Memorial Govt. Medical College Raigarh (CG) without whose support our manuscript would not be a success.

Footnotes

Authors' Contribution: Not mentioned by author.

Conflict of Interests: None declared

Ethical Approval: Not required.

Funding/Support: No funding sources (self-funded).

Informed Consent: Informed consent was taken from the Ethical Committee beforehand.

References

1. Patil R, Palve S, Vell K, Boratne A. Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. *Int J Community Med Public Health*. 2016;**3**:1612–6. doi: 10.18203/2394-6040.ijcmph20161638.
2. van der Vleuten C. Validity of final examinations in undergraduate medical training. *BMJ*. 2000;**321**(7270):1217–9. doi: 10.1136/bmj.321.7270.1217. [PubMed: 11073517]. [PubMed Central: PMC1118966].
3. Javaeed A. Assessment of higher ordered thinking in medical education: Multiple choice questions and modified essay questions. *MedEdPublish*. 2018;**7**(2). doi: 10.15694/mep.2018.0000128.1.
4. Drew S. Perceptions of what helps learn and develop in education. *Teach Higher Educ*. 2010;**6**(3):309–31. doi: 10.1080/13562510120061197.
5. Scouller K. The influence of assessment method on students' learning approaches: Multiple-choice question examination versus assignment essay. *Higher Educ*. 1998;**35**(4):453–72. doi: 10.1023/a:1003196224280.
6. Mehta G, Banode S, Adwal S. Analysis of multiple choice questions (MCQ): Important part of assessment of medical students. *Int J Med Res Rev*. 2016;**4**(2):199–204. doi: 10.17511/ijmrr.2016.i02.013.
7. Premadasa IG. A reappraisal of the use of multiple choice questions. *Med Teach*. 1993;**15**(2-3):237–42. [PubMed: 8246720].

8. Sahoo DP, Singh R. Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students. *Int J Res Med Sci*. 2017;**5**(12). doi: 10.18203/2320-6012.ijrms20175453.
9. Chandra S, Katyal R, Chandra S, Singh K, Singh A, Joshi HS. Medical education/original article creating valid multiple-choice questions (MCQs) Bank with Faculty Development of Pharmacology. *Indian J Physiol Pharmacol*. 2018;**62**(3):359–66.
10. Fozzard N, Pearson A, du Toit E, Naug H, Wen W, Peak IR. Analysis of MCQ and distractor use in a large first year Health Faculty Foundation Program: Assessing the effects of changing from five to four options. *BMC Med Educ*. 2018;**18**(1):252. doi: 10.1186/s12909-018-1346-4. [PubMed: 30404624]. [PubMed Central: PMC6223017].
11. [No author listed]. *Introduction to item analysis. Advancing knowledge, transforming lives*. Michigan State: University trustee group; 2009.
12. Kadiyala S, Gavini S, Kumar DS, Kiranmayi V, Rao PVLNS. Applying blooms taxonomy in framing MCQs: An innovative method for formative assessment in medical students. *J Dr. NTR Univ Health Sci*. 2017;**6**(2). doi: 10.4103/2277-8632.208010.
13. Poornima S, Vinay M. The science of constructing good multiple choice questions (MCQs). *RGUHS J Med Sci*. 2012;**2**(3):141–5.
14. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med*. 1990;**65**(9 Suppl):S63–7. doi: 10.1097/00001888-199009000-00045. [PubMed: 2400509].
15. Shah CJ, Baxi SN, Parmar RD, Parmar D, Tripathi CB. Item analysis of MCQ from presently available MCQ books. *Practising Doct*. 2011;**6**:26–30.
16. Schuwirth LW, van der Vleuten CP, Donkers HH. A closer look at cueing effects in multiple-choice questions. *Med Educ*. 1996;**30**(1):44–9. doi: 10.1111/j.1365-2923.1996.tb00716.x. [PubMed: 8736188].
17. Steinert Y, Mann K, Centeno A, Dolmans D, Spencer J, Gelula M, et al. A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME guide no. 8. *Med Teach*. 2006;**28**(6):497–526. doi: 10.1080/01421590600902976. [PubMed: 17074699].
18. Palmer EJ, Duggan P, Devitt PG, Russell R. The modified essay question: Its exit from the exit examination? *Med Teach*. 2010;**32**(7):e300–7. doi: 10.3109/0142159X.2010.488705. [PubMed: 20653373].
19. Walke Y, Kamat A, Bhounsule S. A retrospective comparative study of multiple choice questions versus short answer questions as assessment tool in evaluating the performance of the students in medical pharmacology. *Int J Basic Clin Pharmacol*. 2014;**3**:1020–3. doi: 10.5455/2319-2003.ijbcp20141212.
20. Azer SA. Assessment in a problem-based learning course: Twelve tips for constructing multiple choice questions that test students' cognitive skills. *Biochem Mol Biol Educ*. 2003;**31**(6):428–34. doi: 10.1002/bmb.2003.494031060288.
21. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ*. 2003;**37**(9):830–7. doi: 10.1046/j.1365-2923.2003.01594.x. [PubMed: 14506816].
22. Chaudhary N, Bhatia BD, Mahato SK. Framing a well-structured single best response multiple choice questions (MCQs)-An art to be learned by a teacher. *J Univers Col Medl Sci*. 2014;**2**(2):60–4.
23. *Rules for writing multiple-choice questions. Annual university conference*. Brigham. Brigham Young University; 2001.
24. FMNHS. *Multiple choice questions guidelines*. Australia: Monash University; 2014. Available from: https://www.monash.edu/__data/assets/pdf_file/0009/1437714/Multiple-Choice-Question-Business-Process.pdf.
25. Kasule OH. Overview of medical student assessment: Why, what, who, and how. *J Taibah Univ Med Sci*. 2013;**8**(2):72–9. doi: 10.1016/j.jtumed.2013.07.001.
26. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Med Educ*. 2008;**42**(2):198–206. doi: 10.1111/j.1365-2923.2007.02957.x. [PubMed: 18230093].
27. Ware J, Vik T. Quality assurance of item writing: during the introduction of multiple choice questions in medicine for high stakes examinations. *Med Teach*. 2009;**31**(3):238–43. doi: 10.1080/01421590802155597. [PubMed: 18825568].
28. Ben-David MF. AMEE guide no. 18: Standard setting in student assessment. *Med Teach*. 2000;**22**(2):120–30. doi: 10.1080/01421590078526.
29. Halikar S, Godbole V, Chaudhari S. Item analysis to assess quality of MCQs. *Med Sci*. 2016;**6**:123–5.
30. Namdeo S, Sahoo B. Item analysis of multiple choice questions from an assessment of medical students in Bhubaneswar, India. *Int J Res Med Sci*. 2016;**4**:1716–9. doi: 10.18203/2320-6012.ijrms20161256.
31. Gajjar S, Sharma R, Kumar P, Rana M. Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian J Community Med*. 2014;**39**(1):17–20. doi: 10.4103/0970-0218.126347. [PubMed: 24696535]. [PubMed Central: PMC3968575].
32. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *J Pak Med Assoc*. 2012;**62**(2):142–7. [PubMed: 22755376].
33. 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 Singapore*. 2006;**35**(2):67–71. [PubMed: 16565756].
34. Vyas R, Supe A. Multiple choice questions: A literature review on the optimal number of options. *Natl Med J India*. 2008;**21**(3):130–3. [PubMed: 19004145].