Indian J Physiol Pharmacol 2018; 62(3) : 359–366

Medical Education / Original Article

# Creating Valid Multiple-Choice Questions (MCQs) Bank with Faculty Development of Pharmacology

**Shalini Chandra[1]*, Rashmi Katyal[2], Sameer Chandra[3], Kashmir Singh[3], Arun Singh[2] and H. S. Joshi[2]**

Departments of [1]Pharmacology, [2]Surgery and [3]Community Medicine,
Rohilkhand Medical College (RMCH),
Bareilly (U.P.)

## Abstract

Formulating good quality multiple choice questions (MCQs) is a demanding assignment, especially devoid of faculty development. Precisely constituted MCQs and item statistics are able to judge students' advanced cognitive domains and allied with enhanced grading preciseness and reliability. The study aims toevaluate impact of faculty development program (FDP) on creation of good quality single best MCQ and on the process of item analysis.

**Methods and Material**: This was quasi-experimental, pre-post test design, interventional study.FDP conducted for the 11 faculty members of Pharmacology department.Each item was analyzed for difficulty index (Df I), discrimination index (DI) and distractor effectiveness (DE), item inscription flaw and cognitive level, both pre and post training. Kirkpatrick's four level model was employed to evaluate participants satisfaction level (level1), learning (level 2) and behavior change (level 3) after workshop.

Chi square test, paired t-test were applied and effect size calculated by Cohen's d test

**Results:** All participants were satisfied with FDP. Their knowledge and skills were enhanced with a significant increase noted in learning (level 2: p=0.001:effect size=0.99) and mean scores of MCQ quality (level 3; p=0.001; effect size =0.735) in post-training test. Significant improvement in all indices were reported after FDP [Df I (p=0.001), DI (p=0.02) and DE (p<0.0001)]

**Conclusions:** There is significant improvement in the quality of MCQs constructed by faculty following FDP. A greater number of higher cognitive MCQs were reported and majority MCQs were found within acceptable and recommended standards for construction of MCQs. FDP is necessary for acquisition of skills for MCQ construction and item analysis.

***Corresponding author :**

Dr. Shalini Chandra, Professor, Department of Pharmacology, Rohilkhand Medical College and Hospital, Bareilly, UP. Email id: shalini1974sameer@gmail.com

(Received on December 2, 2017)

## Introduction

Multiple-choice questions (MCQs) are one of the popular and accepted means of evaluation in medical education. MCQ test items are advantageous as they can cover wider section of lessons and scrutinize large numbers of students in lesser time simultaneously. The tests can be employed for both paradigms of assessment (formative and summative). Colleges are incorporating MCQs tests in their examinations as there is rising trend of adopting MCQs for postgraduate medical entrance examinations. Its acceptance is based on its objectivity, feasibility, high internal consistency and accuracy.

Good quality MCQs compel students to apply advanced level of cognitive processing rather recalling the solitary information (1). Numerous studies have reported that the quality of MCQ test items that medical institutes built are often of poor quality (2). Creating good MCQs are time consuming, challenging and difficult to construct but are effortlessly and consistently scored (3). Characteristics of good quality MCQs are mentioned in terms of item, the stem, and the distractors. MCQs having imperfect stem and unconvincing distractors hinder accurate assessment (4).

Item analysis provides quantitative data at the item-level through knowledge about item statistics and is used extensively to improve test value (5). Difficulty index or facility value (Df I), Discrimination index (DI) and distractor (DE) are the foremost item statistics relevant for estimating the quality of MCQs (6) Abundance of MCQ books of diverse subjects are existing in the market. Many of us either built up test items by ourselves or trusted on questions specified in these books. Nonetheless Shah et al (7) in scrutiny of such MCQs taken straight from such books reported that these MCQs might not be of better-quality and need to be to be analyzed prior to their use for assessment.

Medical faculties often perform those duties for which no formal training has been received by them. Item development and its analysis are one of the duties in which they have no experience and training. There is more possibility of error in framing items if their staff members are not well sensitized and professionally trained enough for the development of test items (8) which leads to lack of quality of many examinations.

Steinert et al. (9) in their systematic review in health professions education reported that faculty training were found to be associated positively with teaching effectiveness both immediate and long term. FDPs for that reason is crucial for development of valid and reliable assessment material. Very few researches have been carried out for faculty development particularly for item writing and it's analysis in India.

We at our department have been executing MCQ based internal exams for undergraduate medical students but faculty members are not trained either in item writing or item analysis. Hence this study was undertaken with the aim of creating valid MCQs bank with the faculty development and the objectives to be achieved were assessment of the impact of FDP and overall satisfaction of the participants. This was the first effort of this kind at our institute.

## Methodology

The study was approved by Ethical committee of the institution. The aim was to assess the impact of training on developing high quality single best MCQs and on the process of item analysis. The target population was eleven faculty members of department of Pharmacology. The secondary purpose was to evaluate the perception on these training workshops. We made use of Kirkpatrick's level of evaluation which comprised of four levels: the trainee's reaction or level of satisfaction with the workshop ( level 1), their learning (level 2), their behavior change (level 3) and long term impact of workshop (level 4) (10).

### Study design

This was a quasi experimental, pre-post test design, interventional study. Preceding to the set up of study the validated faculty satisfaction questionnaire were developed (alpha =0.7196) as per AIMME (11). Eight pre and post short answer questions were developed

regarding MCQ creation and item analysis which were validated by the medical education unit of the institution.

We used structured checklists (12) to review the quality of MCQ items before and after the interventions to reduce the subjectivity while assessing test items. The checklist consisted of 21 markers for assessing MCQ scores. Each marker was allotted one mark thus total marks came out to be twenty one. Each item was scored according to the checklist and scores were calculated out of ten by diving the scores obtained by twenty one and then multiplied by ten. Interventions on three successive days for a period of 3 hours were planned in the department during this project. Before the interventions, departmental meeting was organized and all the faculties were requested to select 30 MCQs on the topic Autonomic nervous system (ANS), what they considered as the best MCQs given in the exam examination with common consensus.

**Day 1:** Pretest (8 short answer questions SAQs) was administered to the participants to assess their knowledge, attitude and perception regarding MCQ framing and item analysis. This was followed by session on MCQ item developing guidelines by imparting theoretical background and all the participants were taught to construct good quality single best answer MCQs.

**Day 2**: Participants were divided into 2 groups. Each group was requested to frame five single best MCQs for practice, and checklist was provided to both groups followed by demonstration by each group which was subsequently peer reviewed by second group and feedback was given.

**Day 3**: Session on item analysis was organized via short interactive lecture. Subsequently participants were requested to fill up the faculty perception questionnaire to evaluate level 1 and post-test was administered to assess learning (level 2)

Following training, the participants were again requested to frame at least 5 MCQs each based on the guidelines discussed in the workshop and submit a total of 30 new MCQs on ANS with final agreement among themselves. Consequently sample of 30 MCQs each before and after the workshop were obtained and analyzed for item quality and item statistics. The results were measured in terms of items of recall and higher cognitive domain, stem flaw, Df I, DI and DE or NFDs.

For interpretation MCQs test paper was administered to second professional M.B.B.S students. Students were divided into two groups. All odd roll numbers are allocated in group A (n=100) and even roll numbers in group B (n=100). Group A was administered 30 MCQs (pre-training) which were submitted by the participants earlier attending workshop .Group B was administered 30 recently framed MCQs (post-training).

Each MCQ comprised of a stem and four choices (one key and three distractors) and students were to opt for one best answer from these four alternatives. Each right answer was allotted one mark and no negative marking for incorrect answer. The answer sheets were collected and grouped according to merit in descending order into three groups. The first group was from the top of the merit (high achiever group, 33%) and third group was bottom of the students (low achiever group, 33%,). The middle 33% were excluded and not used in the item analysis.

**Working Definition formula**

(i) Difficulty Index (Df I) or facility value (FV) : It is calculated using the formula :

$$FV = (HAG + LAG) \times 100 / N$$

HAG = number of students answering the item correctly in the high achiever group, LAG = number of students answering the item correctly in the low achiever group, N = Total number of students

(ii) Discrimination index (DI): This index signifies the ability of a question to discriminate between a higher and a lower achiever student. This is calculated by applying the formula: DI = 2 X (HAG - LAG)/N where the symbols HAG, LAG and N represent the same values as mentioned above. In general, the recommended DI value is >0.25

and DI value 0.15-0.25 is acceptable with revision whereas DI value <0.15 is discarded (13).

(iii) Distractor efficiency (DE): Non Functional Distractor (NFD) in an item is the choice, other than the key selected by less than 5% of students and functional distractor is the option chosen by 5% or more students. On the basis of number of NFDs in an item, DE ranges from 0 to 100%. If an item contained three or two or one or nil NFDs then DE would be 0, 33.3%, 66.6% and 100% respectively (14).

**Statistical analysis**

The data was analyzed in microsoft excel. Chi square and paired t-test were used to evaluate the level of significance. The statistical significance value was specified as $p<0.05$ all through complete analysis. Cohen's d test was adopted for calculating effect size to compare scores obtained in pre and post tests and MCQS ratings pre and post- training.

## Results

All the faculties participated the workshop. The FDP was evaluated according to Kirkpatrick's level of evaluation up to level 3.

Level 1 (reaction) : All the participants filled the faculty satisfaction questionairre .On an average all the participants were satisfied (rating 4-5) with the FDP on a Likert's scale of 1-5. The faculty perception questionnaire and their responses are shown in bar diagram (Fig. 1).

Level 2 (learning): Pre-test and post-test responses were analyzed to assess this level. 8 SAQs were administered both pre and post tests. Mean pre-test score was 2 whereas mean post-test score was 9.38 ($p=0.001$). The effect size was 0.99 which was large according to Cohen's classification (0.2=small; 0.5=medium and 0.8=large). This reflects significant learning (Table I).

Level 3 (behavior change or transfer): This level was assessed by comparing scores of MCQs and different indices of item analysis pre-training versus post-training.. The results are tabulated below from Table II to Table IV. Flow chart showing evaluation across Kirkpatrick's level is displayed in Fig. 2.

Table I shows mean scores and effect size of MCQs to demonstrate improvement in quality of MCQS as per checklist scoring pre-training versus post-training. The quality of MCQS before training was low (6.06±1.225) but higher in post-training (8.07±0.465) with effect size of 0.735 which signified medium to

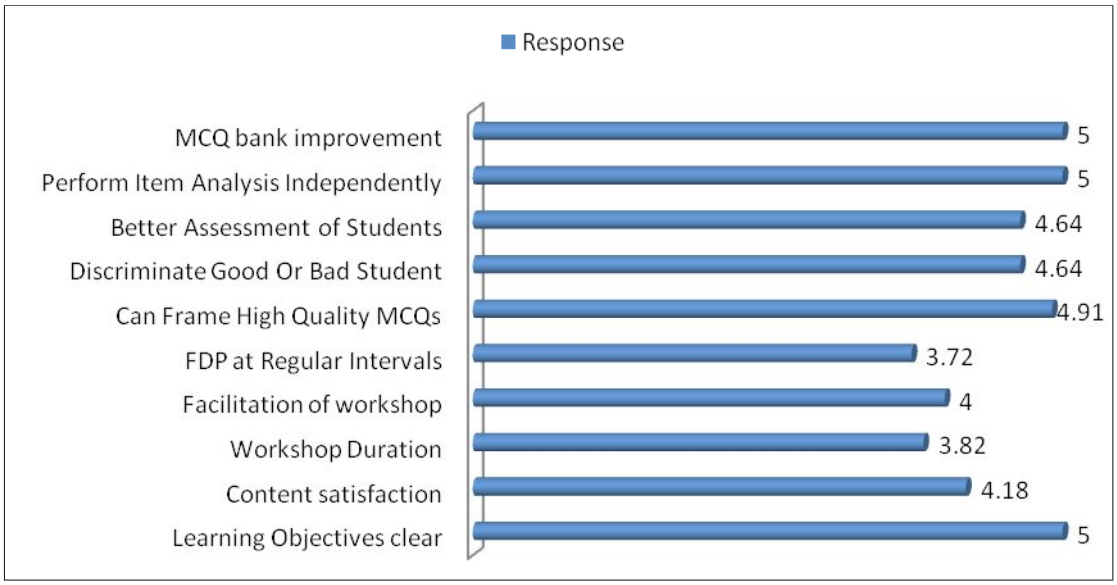


Fig. 1 : Showing faculty perception.

TABLE I: Participants' pre-test and post–test scores to assess learning and mean MCQ scores as per checklist to assess quality of MCQs pre and post training.

| *Number of participants N=11* | *Pre-test and post–test scores to assess learning* | | | |
|---|---|---|---|---|
| | *Pre-test mean* | *Post-test mean* | *P value* | *Effect size* |
| | 2.00±0.632 | 9.36±0.203 | 0.001* | 0.99 |
| | *Mean MCQ scores as per checklist* | | | |
| | *Mean MCQ score Pre-training* | *Mean MCQ score Post-training* | *P value* | *Effect size* |
| | 6.06±1.225 | 8.07±0.465 | 0.001* | 0.735 |

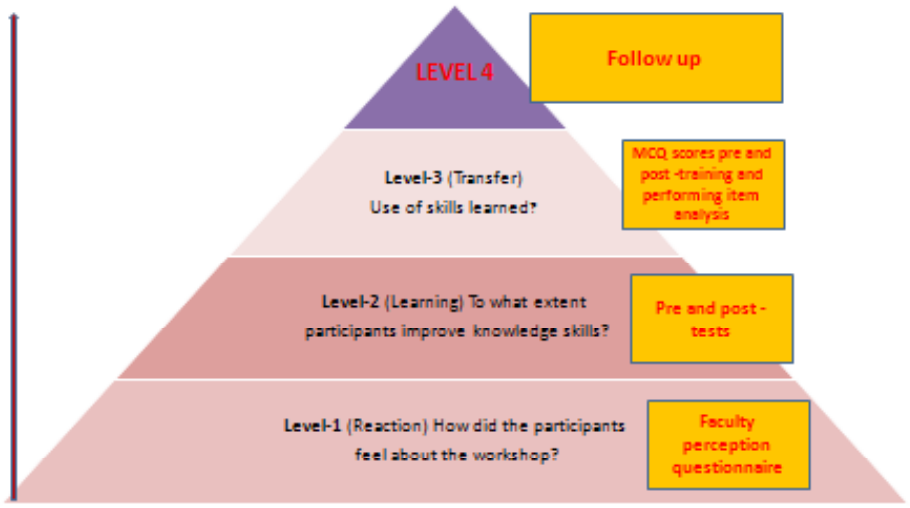


Fig. 2: Showing flow chart showing evaluation across Kirkpatrick's levels of evaluation.

large effect size.

Table II shows that the number of moderate MCQS were more (n=23) in post-training as compared to pre-training (n=11). The number of easy and difficult MCQs were reduced. There is overall improvement in the Df I with significant p value (chi square=8.93; p=0.01). There is increase in number of MCQs having recommended DI (n=12) post-training as compared to pre-training (n=3). Discarded and negative DI MCQs have been reduced (n=6 and n=1 respectively) post-training than pre-training (n=12 and n=6 respectively). The overall DI showed significant improvement (chi square=7.73; p=0.02). Table 2 also depicts that the number of NFDs were also reduced post-training and MCQs having 100% DE were increased to 25(84%) as compared to 6 (20%) and there is highly significant improvement in DE (chi Square=35; 9<0.001).

Table III shows the matrix of the number of items falling in different ranges of difficulty versus discrimination indices both pre and post-training. We found that in post-training 15 out of 30 items were such that they were in the acceptable ranges of both Df I and DI as compared to pre-training in which only 5 out of 30 items were in acceptable range of both DfI and DI . All other items need revision either for Df I and DI.

Table IV shows post-training higher domains MCQs were more (n=22) as compared to pre-training(n=6), whereas recall MCQs were reduced from 24 to 8 after training (chi square=17.14; p<0.001). The MCQs

TABLE II : Ranges of difficulty index, discrimination index and distractor efficiencypre-training vs post-training.

| *Difficulty index* | *Pre-training (n=30)* | *Post-training (n=30)* |
|---|---|---|
| <30 (difficult) | 8 (26.66%) | 5 (16.66%) |
| 30-70 (moderate) | 11 (36.66%) | 22 (73.33%) |
| >70 (easy) | 11 (36.66%) | 3 (10%) |
| Chi square ($\chi^2$=8.93 ); p=0.01* | | |
| *Discrimination index* | *Pre-training (n=30)* | *Post-training (n=30)* |
| >.25(Recommended) | 3 (10%) | 12 (40%) |
| .15-.25(Acceptable) | 15 (50%) | 12 (40%) |
| <.15 (Discarded) | 7 (23.3%) | 5 (16.7%) |
| Negative | 5 (16%) | 1 (3.33%) |
| Chi square ($\chi^2$=8.73); p=0.03* | | |
| *Distractor efficiency* | *Pre-training (n=30)* | *Post-training (n=30)* |
| 100% (NFD=0) | 6 (20%) | 25 (84%) |
| 66.66% (NFD=1) | 10 (33.3%) | 4 (13%) |
| 33.33% (NFD=2) | 8 (26.7%) | 1 (3%) |
| 0% (NFD=3) | 6 (20%) | 0 (0%) |
| Chi square ($\chi^2$=35.46); p=0.000003515* | | |

TABLE III : Distribution of items over the range of Df I and DI.

| *Discrimination index (DI)* | *(Pre-training) (n=30) Difficulty index (Df I)* | | | |
|---|---|---|---|---|
| | *<30* | *30-70* | *>70* | *Total* |
| >0.2 | 0 | 5 | 1 | 6 |
| 0.15-0.2 | 6 | 2 | 4 | 12 |
| <0.15 | 1 | 2 | 4 | 7 |
| Negative | 1 | 2 | 2 | 5 |
| Total | 8 | 11 | 11 | 30 |
| *Discrimination index (DI)* | *(Post-training) (n=30) Difficulty index (Df I)* | | | |
| | *<30* | *30-70* | *>70* | *Total* |
| >0.2 | 0 | 15 | 1 | 16 |
| 0.15-0.2 | 4 | 3 | 1 | 8 |
| <0.15 | 2 | 3 | 0 | 5 |
| Negative | 0 | 1 | 0 | 1 |
| Total | 6 | 22 | 2 | 30 |

having stem flaw were also decreased after training but statistically non-significant (chi square=1.639; p=0.2008).

TABLE IV : Different characteristics associated with items.

| *Different characteristics associated with Items* | *Pre-training (n=30)* | *Post-training (n=30)* |
|---|---|---|
| *Items as per Bloom's cognitive value* | | |
| Recall | 24 | 8 |
| Higher cognitive domains | 6 | 22 |
| *Chi square ($\chi^2$=17.14); p=0.00003* | | |
| *Stem Flaw* | | |
| Unfocussed stem | 8 | 1 |
| Unnecessary information | 5 | 3 |
| Chi square ($\chi^2$=1.639); p=0.2008 | | |

## Discussion

In the present study FDP was evaluated utilizing Kirkpatrick's model of outcome. Participants were satisfied with the workshop. Although satisfaction level is very preliminary for any evaluation but it is must for any positive change to occur (i.e. long term impact) (15). Many studies (9, 16) also reported that participants were satisfied with the FDPs as it was found to be highly useful and of much relevance (17).

Having framed MCQs according to the guidelines, it is important to analyze the quality of these items and whether these are able to discriminate high and low ability students (18).

In the study, knowledge of the participants improved significantly after training, as mean post-test scores were higher than mean pre-test scores (p=0.001 and effect size=0.99). Our results are coherent with the other studies (9, 16, 19).

The study also noted that FDP significantly improved the overall quality of MCQ items. Several studies (2, 8, 12) are in line with our observations.

It is important to ensure reliability of the test items (20). Classical test theory (CT) item analysis is one of the most common method to calculate the reliability of the test item (21). In this study both DF I (chi square=8.93; p=0.01) and DI (chi square=7.73; p=0.02) improved significantly after training. Jozefowicz et. al. (2) reported that trained faculties

had higher mean scores in comparison to untrained faculties when they drafted the United State Medical Licensing Examination (USMLE) Step-1 questions. Items which discriminate poorly or having high and low Df I should be reviewed by content experts as it decreases the validity of the test (22).

DI gives us information about items which effectively discriminated students who were of higher ability or who were of lower ability. Higher the DI more it an effectively discriminate. In our study 16% of the pre-training MCQs were of negative DI while post-training it decreased to 3.33%. Few studies reported negative DI in 20% and 4% of the test items (11, 23). The explanation of negative DI might be ambiguous framing of questions, wrong selection of distractors ,besides poor preparation of students (21). Items having negative DI must be removed from the question bank as it decreases the validity of the test.

When we analyzed distribution of items over the range of Df I and DI, it was observed that maximum discrimination occurred (n=12, DI>0.25) occured with the acceptable difficulty level (30%-70%) in post-training MCQ items. While in pre-training MCQ items only 5 items discriminated maximally between acceptable range of Df I. Si Mui Sim et.al (24) also supported our observation that maximum discrimination had occurred between 40-70% Df I.

Teachers often spend much time and concentration in framing stem than choosing plausible distractors. However the most difficult task in framing MCQs is selecting appropriate distractors other than the answer key. With the help of distractor analysis we can easily identify the student's responses to different options and any NFD should be removed, revised or replaced from the item (25). In this study the percentage of NFDs in pre-training MCQ was 80% which reduced to 16% post-training. Items having zero NFD were 20% pre-training which increased to 84% post-training. Items with three NFDs were decreased to 0% from 20% post-training. Our findings are supported by the results by Abgulghani et al (8).

We also analyzed the quality of MCQs in terms of cognitive level and item writing flaws. Because of simplicity we followed the taxonomy by Tarrant et.al (26). In the current study, higher cognitive domain MCQs (22) were more post-training in comparison to pre-training (8) whereas recall type MCQs were reduced to 6 from 24 after training (chi square=17.14; $p<0.001$; statistically significant). The MCQs with item writing flaws (IWFs) were reduced to 4 from 13 (chi square=1.639; $p=0.2008$; however statistically non significant. Our findings are supported by a number of studies (8, 12, 18). Vyas and Supe (27) stated that lack of faculty training and less time devoted by faculty in MCQ framing mainly contributed to item writing flaws. In the present study we also scored the MCQs pre-training against post-training with structured checklist. There was significant increase in the mean scores of MCQs subsequent to training ($p=0.001$and effect size=0.74). Former researches supported our observations (12, 28, 29).

#### Limitations

One limitation of the study is the smaller sample size. Another limitation is that the FDP was focussed mainly on MCQs, so future workshop is required for other assessment tools. Internal consistency was also not calculated. Moreover, long term impact (level 4) of FDP was not assessed.

#### Conclusion

FDP had significantly improved faculty's competence to develop valid MCQs. Kirkpatrick's model of evaluation provided an effective framework for the FDP. However further research is required to interpret long term impact of the program.

## References

1. Kolte V. Item analysis of multiple choice questions in physiology examination. *IJBAMR* 2015; 4: 320–326.
2. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. *Acad Med* 2002; 77: 156–161.
3. Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. *Educ Psycol Meas* 2004; 64: 391–418.

4. Collins J. Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics* 2006; 26: 543–551.
5. Sayyah M, Vakili Z, Alavi N M, Bigdeli M. An item analysis of written multiple-choice questions: Kashan University of Medical Sciences. *Nurs Midwifery Stud* 2012; 1: 83–87.
6. Downing, S.M. Item response theory: Applications of modern test theory in medical education. *Med Educ* 2003; 37: 739–745.
7. Shah C, Baxi S, Parmar R, D P, Tripathi C. Item analysis of MCQ from presently available MCQ Books. *I J P D* 2011; 6: 26–30.
8. Abdulghani H M, Ahmad F, Irshad M, Khalil MS, AlShaikh GK, Syed S, *et al*. Faculty development programs improve the quality of multiple choice questions items' writing. *Sci Rep* 2015; 5: 9556.
9. Steinert Y, Mann K, Centeno A, Dolmas 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: 497–526.
10. Kirkpatrick DL, Kirkpatrick JD. Evaluating training programs: the four levels. 3rd ed. San Francisco CA: Berrett-Koehler publishers, 2006.
11. Anthony R. A, Jeffrey S. La R, Kent J. D, Hunter G. Developing questionnaires for educational research: AMEE Guide NO. 87. *Med Teach* 2014; 36: 463–474.
12. Naeem N, Vleuten Cees V D, Alfaris E A. Adv in health. *Sci Educ* 2012; 17: 369–376.
13. Singh T, Gupta P, Singh D. Test and item analysis in principles of medical education. Fourth ed. New Delhi Jaypee Brothers Medical Publishers (P) Ltd; 2013; p. 108–113.
14. 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 Ahemdabad, Gujarat. *Indian J Community Medicine* 2014; 39: 17–20.
15. Belfield C, Thomas H, Bullock A, Eynon R, Wall D. Measuring effectiveness for best evidence medical education: a discussion. *Med Teach* 2001; 23: 164–170.
16. AlFaris E, Naeem N, Irfan F, Qureshi R, Saad. H, Sadhan R. A, *et al*. A One-day dental faculty workshop in writing multiple-choice questions: An Impact Evaluation. *J Dent Educ* 2015; 79: 1305–1313.
17. Iramaneerat C. The impact of item writer training on item statistics of multiple-choice items for medical student examination. *Siriraj Med J* 2012; 64: 6.
18. 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. *IeJSME* 2009: 3: 2–7.
19. Markert RJ, O'Neill SC, Bhatia SC. Using a quasi-experimental research design to assess knowledge in continuing medical education programs. *J Contin Educ Health Prof* 2003; 23: 157–161.
20. Senanayake MP, Mettananda DSG. Standards medical students set for themselves when preparing for the final MBBS examination. *Annals Acad Med* 2005; 34: 483–485.
21. Zubairi AM, Kassim NLA. Classical and Rasch analysis of dichotomously scored reading comprehension test items. *Malaysian J of ELT Res* 2006; 2: 1–20.
22. Meshkani Z, Abadie H. Multivariate analysis of factors influencing reliability of teacher made tests. *Journal of Med Ed Winter* 2005; 6: 149–159.
23. Hingorjo MR, Laleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *J Pak Med Assoc* 2012; 62: 142–147.
24. Si-Mui Sim, 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; 67–71.
25. Mehta G, Mokhassi V. Item Analysis of multiple choice questions-an assessment of assessment tool: *Intern J of Health Sci & Res* 2014; 4: 197–202.
26. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple choice questions used in high stakes nursing assessments. *Nurse Educ Today* 2006; 26: 662–671.
27. Vyas R., Supe A. Multiple choice questions: a literature review on the optimal number of options. *Natl Med J India* 2008; 21: 130–33.
28. Kim J, Chii Y, Huensch A, Jun H, Li H, Roullion V. A case study on an item writing process:Use of test specifications,nature of group dynamics, and individual item writers' charecteristics. *Language Assessment Quarterly* 2010; 7: 156–161.
29. Meyari A, Biglarkhani M, Zandi M, Vahedi M, Miresmaeili A. The effect of education on improvent of design of multiple-choice questions in annual residency exams of dental school. *Iran J Med Educ* 2012; 12: 36–45.