

Will a Short Training Session Improve Multiple-Choice Item-Writing Quality by Dental School Faculty? A Pilot Study

Mark A. Dellenges, DDS, MA; Donald A. Curtis, DMD

Abstract: Faculty members are expected to write high-quality multiple-choice questions (MCQs) in order to accurately assess dental students' achievement. However, most dental school faculty members are not trained to write MCQs. Extensive faculty development programs have been used to help educators write better test items. The aim of this pilot study was to determine if a short workshop would result in improved MCQ item-writing by dental school faculty at one U.S. dental school. A total of 24 dental school faculty members who had previously written MCQs were randomized into a no-intervention group and an intervention group in 2015. Six previously written MCQs were randomly selected from each of the faculty members and given an item quality score. The intervention group participated in a training session of one-hour duration that focused on reviewing standard item-writing guidelines to improve in-house MCQs. The no-intervention group did not receive any training but did receive encouragement and an explanation of why good MCQ writing was important. The faculty members were then asked to revise their previously written questions, and these were given an item quality score. The item quality scores for each faculty member were averaged, and the difference from pre-training to post-training scores was evaluated. The results showed a significant difference between pre-training and post-training MCQ difference scores for the intervention group ($p=0.04$). This pilot study provides evidence that the training session of short duration was effective in improving the quality of in-house MCQs.

Dr. Dellenges is HS Clinical Professor, Department of Preventive and Restorative Dental Sciences, School of Dentistry, University of California, San Francisco; and Dr. Curtis is Professor, Department of Preventive and Restorative Dental Sciences, School of Dentistry, University of California, San Francisco. Direct correspondence to Dr. Mark Dellenges, Department of Preventive and Restorative Dental Sciences, D-3214, University of California, San Francisco, School of Dentistry, 707 Parnassus Avenue, San Francisco, CA 94143-0758; 415-476-1785; mark.dellenges@ucsf.edu.

Keywords: dental education, dental faculty, assessment, faculty development, multiple-choice questions, item-writing flaws

Submitted for publication 11/11/16; accepted 1/3/17
doi: 10.21815/JDE.017.047

Multiple-choice questions (MCQs) are a useful testing format for evaluating cognitive knowledge and are frequently used in health professions education.¹ MCQs can be used to assess factual recall, comprehension, and application,² and they have advantages of being reliable and versatile yet requiring careful item development to ensure fairness and validity.^{3,4} Most examinations are developed in-house by faculty members who teach the courses, but few faculty members have received training in developing high-quality MCQs.⁵ Two studies found that extensive and repeated faculty development programs aimed at test development improved in-house item-writing by medical school faculty.^{6,7}

Without specific training, most novice item writers tend to create poor-quality, flawed, low-cognitive-level test questions that test unimportant or trivial content.⁸ Flawed items can result if standard item-writing guidelines or principles are not adhered to when writing items. Test-item construction manuals like those by Gronlund as well as Case and

Swanson outline item-writing guidelines.^{9,10} Guidelines for writing MCQs have also been developed and revised to focus on categories of content, format, style, writing the stem, and writing the options.¹¹⁻¹⁴ For instance, MCQs with negatively worded stems or the use of "all of the above" or "none of the above" in the options are considered flawed.¹⁴ According to standard item-writing principles, negatively phrased stems and use of "all of the above" or "none of the above" in the options should be avoided.

Violations of the most basic item-writing principles are common in achievement tests used in health professions education.¹⁵⁻¹⁹ In a medical basic science achievement test, Downing found that 11 out of 33 questions (33%) were flawed.²⁰ In assessing the quality of four examinations given in a U.S. medical school, Downing found that 46% of the MCQs contained item-writing flaws.¹⁶ In another study, test-writing errors by faculty members in nursing education were found to occur in 46.2% of the 2,270 MCQs collected from tests and examinations over a five-year period from 2001 to 2005.⁵

Even in high-stakes nursing achievement tests, the same study found an average of 47.3% of all MCQs were flawed. Negative stems, unfocused stems, and “window dressing” (i.e., excessive verbiage) were the most frequently observed item flaws in Tarrant and Ware’s study.¹⁷ Another study reported that 85% of MCQs had at least one flaw in a hospital professional development program for nurses.¹⁸

Tests containing flawed or poorly written items can have consequences for the test-takers. Downing applied established principles of effective multiple-choice item-writing to find flawed and unflawed items and then compared reliabilities and item difficulties.¹⁶ The flawed items reduced reliability coefficient values from 0.62 to 0.44 after correcting for test length while being more difficult and less discriminating. In that study, 10-15% of students’ tests that were classified as failures would have been passes if items with questions with item-writing flaws were removed. Tarrant and Ware reported that, among nursing students, flawed items adversely affected higher achieving students more than lower achieving students.¹⁷

Multiple-choice items that fail to adhere to evidence-based guidelines may introduce construct-irrelevant variance to an assessment. Construct-irrelevant variance (CIV) alters test score variation by introducing factors unrelated to the measurement of the intended construct. CIV may impact the difficulty of a test question, independent of the content of the MCQ, and can result in erroneous test scores and pass-fail decisions.²⁰ Besides the use of poorly crafted or flawed test items, cheating or unsecure test questions can contribute to CIV. Construct-irrelevant variance can also be introduced by student guessing, “test-wiseness” (defined as any skill that allows a student to choose the correct answer on an item without knowing the correct answer), and item bias such as differential item functioning or using indefensible passing scores.^{1,20}

Faculty development programs have shown value in improving quality of test items, yet most programs have involved a significant faculty time commitment.^{5,15-17,21-24} Jozefowicz et al. conducted one of the rare studies of faculty development programs in the health professions that focus on improvement of test development and standard setting.¹⁹ They found that untrained test item writers were not as effective at writing exam items as those trained using a standard method, such as the one outlined in the National Board of Medical Examiners text on item-writing, *Constructing Written Test Questions for the Basic and Clinical Sciences*.^{10,19,24}

A study by Iramaneert suggested that faculty development workshops of varying length over extended periods of time can improve test-writing skills of faculty members.⁷ The first workshop in that study consisted of a three-hour session on MCQ item development, followed by two additional two-hour workshops three months later. The content included classical item analysis, with the analysis used as feedback to improve item-writing. Item difficulty and item discrimination statistics from comprehensive examinations given one year prior to the workshops (pre-training) and six months after the workshops (post-training) were analyzed to determine the impact of the workshops on item quality. The results suggest that the workshops were effective in improving the quality of test items as demonstrated by the improvement of post-training item difficulty and discrimination indices. In another study, Naeem et al. described an even more time-intensive one-week faculty development program in which faculty members were instructed to write MCQs, short-answer questions, and checklists for an objective structured clinical examination (OSCE).⁶ To evaluate the effects of the program, the authors asked participants to submit an example of their “best” item for each of the item categories prior to the start of the program. Participants then rewrote their test items after each phase of the intervention. The test items were scored at pre-training, at midpoint, and after the second intervention. There was a significant increase in scores from pre-training to mid-point assessment and from mid-point to post-training with strong effect sizes. The results of these two studies provide evidence that the quality of test items can improve through faculty training.

Though effective, those previous multiple-session and weeklong, full-time item-writing workshops were resource-intensive and required a significant amount of faculty time. The aim of our pilot study was to determine if a short workshop would result in improved MCQ item-writing by dental school faculty. We hypothesized that having faculty members participate in a one-hour training session on improving MCQ item-writing skills (intervention) would improve the quality of their MCQs.

Methods

The study was reviewed and exempted from Institutional Review Board review by the University of California, San Francisco (UCSF), Human

Research Protection Program because it involved a commonly accepted educational practice in an established educational setting and involved only data obtained from standard educational tests. At the UCSF School of Dentistry, third-year dental students are required to take a four-quarter didactic clinical general dentistry course that meets for four hours per week. This patient-centered care course is made up of 13 modules covering various topics and disciplines in general dentistry. At least 55 faculty members from multiple departments in the school provide lectures in the course. The instructors are asked to submit MCQs from their lecture material. Examinations are developed from these MCQs and are used to assess student achievement at the end of each module.

For this pilot study conducted in 2015, 24 faculty members who had provided a minimum of six in-house MCQs in the last three years were chosen to participate. From the MCQs collected, a random sample of six MCQs written by each of the 24 faculty members was obtained using a random number generator in the Stata Statistics and Data Analysis Software Program (Stata Version 13, StatCorp, College Station, TX, USA). The items were each assigned a unique code consisting of item number, faculty identification, and experimental status (assignment to either the control or treatment group). These MCQs became the pre-training set of items for the experimental groups.

The 24 dental school faculty members were divided into two groups: a no-intervention (control) group and an intervention (treatment) group. To control for experimenter bias, 12 faculty members were

randomly selected for the no-intervention group, and the remaining 12 faculty members were assigned to the intervention group.

The faculty members in the intervention group were asked to participate in a one-hour training session to improve their MCQ item-writing. The training session consisted of a 30-minute PowerPoint presentation (Microsoft PowerPoint for Mac 2011, version 14.6.9) on improving MCQ quality, along with discussion of examples of poorly constructed and improved MCQ items provided by a faculty trainer. Faculty members were also provided a handout of the PowerPoint presentation and an MCQ item improvement checklist adapted and modified from Naeem et al.⁶ (Table 1).

The faculty members in the intervention group were given their six previously written, randomly selected pre-training MCQs and were asked to revise them utilizing what they had learned during the training session and then to return revised questions to the authors within five days. These items were also assigned a unique code consisting of item number, faculty identification, and experimental status and became the post-training set of items for the intervention group.

Faculty members in the no-intervention group, who also contributed six pre-training questions, were told that writing high-quality test items is important, but they did not participate in the MCQ training session/discussion or receive the presentation handout or checklist for improving MCQs. They were provided with their six previously written, randomly selected, and coded pre-training items. The faculty members in the no-intervention group were asked to revise

Table 1. Checklist for improving multiple-choice questions (MCQs)

MCQs Checklist for Faculty Training Session

Question is appropriate for student's level.

A single clearly formulated problem in simple language is presented in the stem of the item. As much of the wording as possible is in the stem.

The stem of the item is stated in positive form whenever possible; if negative words are used, they appear capitalized and in bold type.

All options are uniform in grammatical construction and length.

Only one of the options is the correct answer.

Verbal cues are avoided that might enable students to select the correct answer or to eliminate an incorrect alternative. (Absolute words such as "always," "never," "all," "none," or "only" used in the distractors are commonly associated with false statements.)

The option "all of the above" is avoided, and "none of the above" is used only with extreme caution.

Source: Adapted from Naeem N, Van der Vleuten CPM, Alfari EA. Faculty development on item-writing substantially improves item quality. *Adv Health Sci Educ* 2012;17(3):369-76.

their previously written items and to return them to the authors within five days. These items were again assigned a unique code consisting of item number, faculty identification, and experimental status and became the post-training set of items for the no-intervention group.

The pre-training and post-training questions were rated using the modified MCQ scoring rubric by two faculty raters. The raters chosen to rate MCQ item quality held faculty positions at the dental school but were not trained dentists. One rater was a dental hygienist with a doctorate in education, and the other was a staff psychologist at the dental school. The raters were trained and calibrated by independently receiving two hours of training and instructions for rating the item quality of questions using the modified MCQ scoring rubric (Table 2). The raters evaluated the coded pre-training and post-training MCQs and were blinded to which faculty members were in the no-intervention and intervention groups. The maximum score attainable for each item rated with the modified MCQ scoring rubric was seven points. The average item quality scores for the pre-training and post-training set of MCQs between the no-intervention and intervention group were calculated and reported.

Since each faculty member contributed six pre-training and six post-training MCQs, we were unable to use techniques that assume independence of observations such as analysis of variance. For each faculty member, 24 ratings were made for the questions they contributed: six before intervention/control activity and six after, by each of two raters. The data were then clustered based on three categories: within question (before and after), within faculty member (each contributed six questions before and after),

and within rater (two raters each rated all before and after questions). Because data within each cluster were likely to be correlated (e.g., one rater may score questions consistently higher than the other rater), within cluster observations cannot be assumed to be independent of one another.

To accommodate these correlations, we reduced the data as follows: within question correlation was handled by calculating the difference between the before and after ratings for each question rated by the first rater. We then averaged across the six differences calculated for each faculty member to obtain one Rater 1 summary change number per faculty member. This procedure was repeated for the second rater, resulting in one Rater 2 summary change measure for each of the same faculty members rated by Rater 1. Finally, we averaged the two raters' summary change measures for each faculty member; therefore, the unit of analysis was the average change or difference score across six questions for each faculty member, averaged between the two raters. These difference scores, now reduced to independent observations, were then compared between the no-intervention and the intervention groups using a two-sample Wilcoxon rank-sum (Mann-Whitney) test. A measure of interrater reliability was calculated by using the difference scores from both raters to develop a Kappa reliability coefficient.

Results

Eleven of the 12 faculty members in the no-intervention group returned their post-training set of six revised MCQs, and all 12 of the faculty members in the intervention group returned their post-training

Table 2. Sample of modified scoring rubric provided to raters

Rubric Item	Points	Scoring
Question is appropriate for student's level.	1/0 points	overall
Question presents a single clearly formulated problem in simple language in the stem or stem/lead-in of the item; as much of the wording as possible is in the stem or stem/lead-in.	1/0 points	stem or stem/lead-in
Question states the stem of the item in positive form, but if negative words are used, they appear capitalized and in bold type.	1/0 points	stem
Question options are uniform in grammatical construction and length.	1/0 points	options
Only one of the question options is the correct answer.	1/0 points	options
Question avoids verbal cues that might enable students to select the correct answer or to eliminate an incorrect alternative. (Absolute words such as "always," "never," "all," "none," or "only" used in the distractors are commonly associated with false statements.)	1/0 points	options
Question avoids using the alternative "all of the above," and "none of the above" is used with extreme caution.	1/0 points	options

set of six revised MCQs. Interrater reliability between the raters for the pre-training and post-training difference scores yielded a Kappa coefficient of 0.34 with 77% agreement.

The average pre-training item quality score for the no-intervention group was 6.11 (0.61), and the post-training score was 6.16 (0.60). The average pre-training item quality score for the intervention group was 6.13 (0.65), and the post-training score was 6.35 (0.55) (Table 3). The difference scores for the no-intervention and intervention groups are shown in Table 4. The two-sample Wilcoxon rank-sum (Mann-Whitney) test indicated that there was a significant difference between the pre-training and post-training difference scores between the intervention and no-intervention groups ($p=0.04$).

The majority of the faculty members in the no-intervention group showed little to no improvement in MCQ item-writing quality from pre-training to post-training, but there was greater improvement in the intervention group (Figure 1). For the no-intervention group ($N=11$), two faculty members' changes in scores showed improvement in MCQ item-writing quality, while seven faculty members' scores had no change and two faculty members' scores were lower than at pre-training. For the intervention group ($N=12$), eight faculty members' change in scores showed improvement in MCQ item-writing,

while three faculty members' scores had no change and only one faculty member's score was lower than at pre-training.

Discussion

In our experience, most dental schools do not have the resources to provide ongoing training programs for faculty members to improve their MCQ writing skills and rarely take the time for embedded or lengthy faculty development. Our pilot study suggests that significant improvement in MCQ item-writing can occur with a short training session. Shorter sessions are more convenient for faculty, encourage faculty involvement, and are less resource-intensive for dental schools than longer or multiple training sessions. Shorter training sessions could be repeated more often to reinforce faculty retention of the information and could be given at times when instructors are most likely to be writing test items, usually right before examinations.

In this study, the intervention group (who received training) showed a statistically significant improvement in MCQ item quality based on the difference scores between pre-training and post-training. The no-intervention group (who did not receive training) did not show a statistically significant improvement in MCQ item quality.

Our results compare favorably with other studies that have established development programs for faculty to improve MCQ item-writing.^{6,7} The other interventions were of longer duration or had multiple interventions as compared to our single intervention of short duration. Although the intervention group in our pilot study showed statistically significant

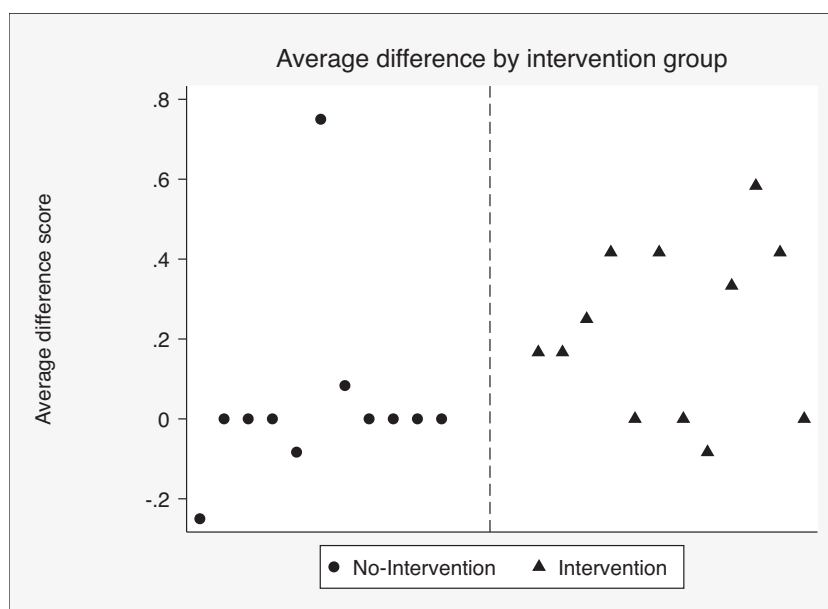
Table 3. Mean item quality scores (standard deviation) for no-intervention and intervention groups at pre- and post-training (score is average out of possible 7)

Group	Pre-Training	Post-Training
No-intervention	6.11 (0.61)	6.16 (0.60)
Intervention	6.13 (0.65)	6.35 (0.55)

Table 4. Difference scores for no-intervention and intervention groups at pre- and post-training

Difference Scores for No-Intervention Group			Difference Scores for Intervention Group		
	Number	Percent		Number	Percent
-0.25	1	9.09%	-0.08	1	8.33%
-0.08	1	9.09%	0.00	3	25.00%
0.00	7	63.64%	0.17	2	16.67%
0.08	1	9.09%	0.25	1	8.33%
0.75	1	9.09%	0.33	1	8.33%
			0.42	3	25.00%
			0.58	1	8.33%
Total	11	100%	Total	12	100%

Note: In the no-intervention group, there were seven faculty scores that, on average, did not change, while two improved and two decreased. In the intervention group, there were three faculty scores that, on average, did not change, while eight improved and only one decreased.



improvement, we were not able to calculate an effect size to compare to what other investigators have found with longer and/or multiple interventions. It is probable that longer or repeated interventions, perhaps just prior to examinations, would be important times for reinforcement of key aspects of writing MCOs.

Limitations of this pilot study include the small number of dental faculty members who participated. Twenty-four faculty members were selected from among those instructors who had written in-house MCQs in the last three years for a large didactic course given to third-year dental students. Only 23 faculty members at one dental school completed the study. Another limitation of the study was the low to moderate Kappa coefficient score. The fact that agreement was not higher likely resulted from the inevitably subjective nature of MCQ evaluation, despite prior training. No training could cover all the potential ways in which a question could be faulty, so raters were required to inject their own opinions to supplement the guidelines we provided. We mitigated this variability by averaging the two raters. Future studies on this topic might benefit by providing standardized MCQs to faculty members

to improve standardization. However, our study had the advantage of using questions created by the participants themselves.

Another possible limitation results from the fact that use of the modified MCQ scoring rubric in this pilot study resulted in pre-training MCQ item quality mean scores that were relatively high to begin with. The post-training MCQ item quality mean scores for the intervention group increased only slightly, from 6.13 to 6.35. Although the difference scores showed that the difference in improvement between the two experimental groups was statistically significant ($p=0.04$), the modest increase in the mean post-training scores for the intervention group may be of questionable practical significance. The modified MCQ scoring rubric used in this pilot study may not have had sufficient sensitivity to discriminate between low and high MCQ item quality attributes. Developing a scoring rubric with an expanded scale may increase the range of scores and provide higher sensitivity. This need could be addressed in future studies. Perhaps reporting the change in the number or quantity of item flaws would provide clearer evidence for improvement of in-house MCQs following intervention.

Another issue to consider is the focus of the study and the construct we set out to measure. Merely showing that a short training session can improve MCQ item-writing quality does not directly prove that examinations made up of the improved MCQs are better or that the dental school faculty will become expert test developers. It is not the intent of faculty development programs to create test development experts. The real intent is to help faculty members recognize and avoid common errors that can adversely affect individual item quality and functioning.

The results of this pilot study may not be generalizable because of the specified limitations. Future research involving a larger sample size, a revised MCQ scoring rubric, and improved inter-rater reliability may result in better reliability and generalizability.

Conclusion

Faculty training involving prolonged and intensive development programs has been shown to improve in-house question-writing quality in medical education. We sought to determine if a one-hour faculty training session could improve in-house MCQ item-writing quality among the faculty at one dental school. The results of this pilot study suggest that a short, one-hour training session for dental school faculty members led to significant improvements in in-house MCQ item-writing quality at one U.S. dental school. Shorter training sessions are more convenient for faculty members and less resource-intensive for dental schools, but can nevertheless result in MCQs that are of improved quality.

Acknowledgments

We are thankful to Dr. Xiaoxia Newton and Dr. Mark Wilson, University of California, Berkeley, as well as Dr. Christy Boscardin, University of California, San Francisco, for their assistance with this manuscript. We are also thankful to Dr. Nancy Hills, University of California, San Francisco, for her guidance with the research statistics. We thank Dr. Linda Centore and Dr. Gwen Essex for their assistance and contributions to this research project.

REFERENCES

1. Downing SM, Yudkowsky R. *Assessment in health professions education*. New York: Routledge, 2009.

2. Albino JE, Young SK, Neumann LM, et al. Assessing dental students' competence: best practice recommendations in the performance assessment literature and investigation of current practices in predoctoral dental education. *J Dent Educ* 2008;72(12):1405-35.
3. Campbell DE. How to write good multiple-choice questions. *J Paediatr Child Health* 2011;47:322-5.
4. Ware J, Torstein VIK. Quality assurance of item-writing during the introduction of multiple-choice questions in medicine for high-stakes examinations. *Med Teach* 2009;31:238-43.
5. 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(8):662-71.
6. Naeem N, Van der Vleuten CPM, Alfari EA. Faculty development on item-writing substantially improves item quality. *Adv Health Sci Educ* 2012;17(3):369-76.
7. Iramaneert C. The impact of item writer training on item statistics of multiple-choice items for medical student examination. *Siriraj Med J* 2012;64(6):178-82.
8. Downing SM, Haladyna TM. *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates, 2006.
9. Gronlund NE. *Assessment of student achievement*. Boston: Allyn & Bacon, 2003.
10. Case SM, Swanson DB. *Constructing written test questions for the basic and clinical sciences*. 3rd ed. Philadelphia: National Board of Medical Examiners, 2003.
11. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. *Appl Meas Educ* 2003;2(1):37-50.
12. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. *Appl Meas Educ* 1989;2(1):51-78.
13. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ* 2002;15(3):309-33.
14. Haladyna TM, Rodriguez MC. *Developing and validating test items*. New York: Routledge, 2013.
15. Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. *Adv Health Sci Educ* 2002;7(3):235-41.
16. Downing SM. The effects of violating standard item-writing principles on tests and students: the consequences of using flawed test items on achievement examination in medical education. *Adv Health Sci Educ* 2005;10(2):133-43.
17. 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.
18. Nedeau-Cayo R, Laughlin D, Rus L, Hall J. Assessment of item-writing flaws in multiple-choice questions. *J Nurses Prof Dev* 2013;29(2):52-7.
19. Jozefowicz RF, Loeppen BM, Case S, et al. The quality of in-house medical school examinations. *Acad Med* 2002;77(2):156-61.
20. Downing SM. Construct-irrelevant variance and flawed test questions: do multiple-choice item-writing principles make any difference? *Acad Med* 2002;77(10):S103-4.

21. Collins J. Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics* 2005;26(2):543-51.
22. Wrigley W, Van der Vleuten CPM, Freeman A, Muijtjens A. A systemic framework for the progress test: strengths, constraints, and issues. AMEE guide no. 71. *Med Teach* 2012;34(9):683-97.
23. Wallach PM, Crespo LM, Holtzman KZ, et al. Use of a committee review process to improve the quality of course examinations. *Adv Health Sci Educ* 2006;11(1):61-8.
24. Hodgson CS, Wilkerson L. Faculty development for teaching improvement. In: Steinert Y, ed. *Faculty development in the health professions: a focus on research and practice*. Dordrecht: Springer, 2014.