Education in Medicine Journal ISSN 2180-1932

ORIGINAL ARTICLE
Volume 6 Issue 2 2014
DOI: 10.5959/eimj.v6i2.216
www.eduimed.com

The effect of MCQ vetting on students' examination performance

Majed Wadi [1], Ahmad Fuad Abdul Rahim[2], Muhamad Saiful Bahri Yusoff[2], Kamarul Aryffin Baharuddin[3]

[1]Medical Education Unit, Faculty of Medicine and Health Sciences, University of Science and Technology, Sana'a, Yemen. [2]Medical Education Department, [3]Emergency Medicine Department, School of Medical Sciences, Universiti Sains Malaysia, Kota Bharu, Kelantan, Malaysia.

ARTICLE INFO

Received : 17/07/2013
Accepted : 23/01/2014
Published : 01/06/2014

KEYWORD
Vetting
MCQ
Assessment

ABSTRACT

Context: Question vetting is important to ensure validity, reliability, and other quality indicators of assessment tools, including the MCQ. Faculty members invest a substantial amount of time and effort into the MCQ vetting process. However, there is shortage of scientific evidence showing its effectiveness and at which level it needs to be focused on. This study aimed to provide scientific evidence regarding the effects of question vetting process on students' examination performance by looking at their scores and pass-fail outcomes. **Method:** A parallel randomized control trial was conducted on third year medical students in a medical school. They were randomly assigned into two equal groups (i.e. control and experimental). Two mock examinations were conducted (i.e. time I and time II). At time I, non-vetted MCQs were administered to both groups as a baseline measurement. At time II, vetted MCQs were administered to the experimental group, while the same non-vetted MCQs were administered to the control group. **Results:** Out of 203 students, 129 (63.5%) participated in both mock examinations. 65 students were in the control group and 64 students were in the experimental group. Statistical analysis showed no significant differences ($p > 0.05$) in mean examination scores and pass-fail outcomes between or within the control and experimental groups. **Conclusion:** This study indicated that the MCQ vetting process did not influence examination performance. Despite these findings, the MCQ vetting process should still be considered an important activity to ensure that test items are developed at the highest quality and standards. However, it can be suggested that such activity can be done at the departmental level rather than at the central level.

CORRESPONDING AUTHOR: Dr Majed Wadi, Medical Education Unit, Faculty of Medicine and Health Sciences, University of Science and Technology, Sana'a, Yemen. The 60th road, P.O.Box: 13064.
Email: majed_wadi@yahoo.com/ m.wadi@ust.edu

Introduction

Question vetting is the process of reviewing and evaluating test items according to specified criteria to detect flaws and to edit them accordingly to improve their quality (1-6). Vetting is not only important in maintaining a high standard of test items but it also helps sustain their validity. Most produced items, even those produced by experienced item writers, are still flawed in some ways (3, 7). Such items are the major threat of assessment (8-10). They are

frequently encountered in many in-house tests. So, once an item is constructed, it should undergo a critical review by a review (or vetting) committee (5, 10). Establishing a vetting committee is strongly recommended (7) and is attributed to the significant improvement of test item quality (5, 11). In vetting sessions, an item is edited to remove flaws and to make it as clear and understandable as possible. Activities in the vetting committee include a) content review - the content of each item is matched with what is intended to be measured (testing blueprint), b) item-writing principles review - items must be ensured to adhere to identified item-writing guidelines, c) editorial review - items are checked for any errors in spelling, grammar, and punctuation, and d) answer key check - each item is checked for accuracy of the correct answer (3).

We found limited studies which evaluated the impact of such a process on examination performance. In one study (8), Downing evaluated the construct-irrelevant variance (CIV) associated with flawed test items with respect to examination difficulty and pass-fail decisions. He found that flawed items were seven percentage points more difficult and failing more students than non-flawed items. He recommended future research with an experimental design in which flawed (i.e. non-vetted) and non-flawed (i.e. vetted) items are randomly assigned to examinees to make powerful comparisons and generalizations.

In another study (12), he examined four examinations to investigate the effects of the violation of multiple-choice item writing principles on test characteristics, student scores and pass-fail outcomes. He found that violated (flawed) items were more difficult (higher failure rate) than standard (unflawed) items.

A similar study was done by Tarrant and Ware (13) to examine the impact of item-writing flaws on student achievement in high-stakes nursing assessments. In contrast to Downing's findings, they found that unflawed items were associated with lower passing and higher failure rate than flawed items. They also noted that standard (unflawed) items scored higher than flawed items.

This study was conducted to provide scientific evidence on the effect of MCQ vetting on student examination performance in the School of Medical Sciences (SMS), Universiti Sains Malaysia (USM) through comparing the mean exam scores and pass-fail outcomes between vetted and non-vetted MCQs.

Method

A single-blinded parallel randomized controlled trial (RCT) study design was utilized. Two mock examinations were conducted at two different times (designated time I and II) with a gap of two months in between. A panel of academic staff from various departments was recruited to provide non-vetted MCQs to be used in the examinations. Three types of MCQ; Multiple True-False (MTF), Single Best Answer (SBA), and Extended Matching Question (EMQ), were chosen. A stratified random sampling method was used to select the MCQs from the generated item pool. A total of 70 items; 50 MTF items, 10 SBA items and 10 EMQ items, were selected. These items covered three blocks: Respiratory (Resp.), Cardiovascular (CVS) and Gastrointestinal (GIT). Anatomy, Physiology, Biochemistry and Pathology subjects were tested. One hour was allocated for each exam. At time I, the non-vetted MCQs were administered to both groups (i.e. experimental and control) as a baseline measurement. Students were not allowed to take back the question paper with them after test I.

The selected MCQs were then vetted by the phase II vetting committee. The committee members vetted the questions based on their past vetting experience. Appendix I contains some MCQs before and after vetting process.

At time II (after two months), vetted MCQs were administrated to the experimental group, while the same non-vetted MCQs were administrated to the control group.

Third-year medical students were recruited as examinees in the mock examinations at time I and II. The students were considered as a medium to reflect the effectiveness of vetting questions in terms of examination mean score and pass-fail outcome. All new third year medical students (excluding those who are repeating the year and those who only attended one of the mock examinations) were invited to take part in this study. There were a total of 203 students and an approximately 20% dropout rate was expected in the both examinations.

A simple randomization was used to assign the medical students into experimental (i.e. vetted) and control (i.e. non-vetted) groups. To ensure equal distribution of students from both groups during the examination, each student was given a number as a student identity code for seating in an examination hall.

Three different optical mark recognition (OMR) answer sheets were administered to both groups in both mock examinations. These were according to MCQ type. Pencils were used to answer MCQ in OMR.

The cut-off point for passing the exam was set as 50%. The student is considered fail if he/she obtained less than 50% and considered pass if he/she obtained above 50%.

Ethical considerations

Ethical approval was obtained from the ethical committee of the School of Medical Sciences, USM. Selected study subjects were assembled and briefed regarding the study purpose, procedures, and examination date and time especially emphasizing that participation will not affect their progression in the course. Study subjects who agreed to participate in the study were required to fill an informed consent form. Each subject was given an identity code (ID) and their confidentiality were preserved and maintained.

Data Analysis

Collected data was cleaned, sorted, and analyzed using different soft-wares as the following:

SmartScan *(14)*

This was used for scoring and analyzing items. The answer sheets were checked for students' number before scanning.

IBM SPSS Statistics 19 *(15)*

This was used for data entry and analysis. Alpha (α) was fixed at 0.05 with a confidence interval of 95%. Assumptions were made before statistical analysis.

STATA 9 *(16)*

This was used to calculate Fisher's exact test when the independent variables are more than two (polytomous). This was applicable when expected cells were equal or less than 5 or more than 20%.

Statistical tests

Independent t/Mann-Whitney test was used to compare mean/median exam scores between groups (control and experimental) at time I and time II. Paired t/Wilcoxon Signed Rank test was used to compare mean/median within groups in both study stages (pre and post). Pearson Chi Square/Fisher's Exact was used to compare pass-fail outcome between groups, while McNemar test was used to compare pass-fail outcomes within groups.

Results

With regards to participation rate, out of 160 medical students that initially participated in the first mock examination, 129 (80.6%) took part in the subsequent mock examination. 65 students (50.4%) were in the control group and 64 students (49.6%) were in the experimental group.

The homogeneity and heterogeneity of the intervention groups are important issues that must be addressed to ensure comparability (17). Suitable statistical tests were run to check homogeneity between the groups. Results showed that the control group had more students with other entry qualification background compared to the experimental group (table 1). This is the solitary difference between the groups.

Table 1: Students homogeneity according to their demographic and academic characteristics in both control and experimental groups

Variable	Control Group (n = 65)	Experimental Group (n = 64)	p-value
Gender, *n* (%)			
Male	17 (26.2)	20 (31.3)	0.552[a]
Female	48 (73.8)	44 (68.8)	
Age			
Mean (SD)	21.17 (0.55)	21.05 (0.60)	0.229[b]
Race, *n* (%)			
Malay	40 (61.5)	32 (50.0)	0.120[c]
Chinese	19 (29.2)	21 (32.8)	
Indian	3 (4.6)	10 (15.6)	
Others	3 (4.6)	1 (1.6)	
Entry qualifications, *n* (%)			
Matriculation	53 (81.5)	58 (90.6)	0.048[c]
STPM	6 (9.2)	6 (9.4)	
Others	6 (9.2)	0 (0)	
Past Academic Records, *n* (%)			
Final Phase I Grades			0.348[c]
A	10 (15.4)	5 (7.8)	
B	14 (21.5)	18 (28.1)	
C	41 (63.1)	40 (62.5)	
F	0	1 (1.6)	
Continuous Assessment I Grades			0.651[c]
A	6 (9.2)	3 (4.7)	
B	13 (20.0)	10 (15.6)	
C	23 (35.4)	25 (39.1)	
F	23 (35.4)	26 (40.6)	
Continuous Assessment II Grades			0.716[c]
A	4 (6.2)	3 (4.7)	
B	12 (18.5)	9 (14.1)	
C	26 (40.0)	23 (35.9)	
F	23 (35.4)	29 (45.3)	

[a]Chi-square, [b] Independent t-test and [c]Fisher's exact tests were applied

Table 2 demonstrates the comparison of mean exam score between the groups (control and experimental) at each stage according to item formats. The mean exam scores across the item formats and stages of examination were not significantly different between the control and experimental groups except for SBA and overall score at the post-stage where the control group scored higher than the experimental group. In general, it appeared that there are no obvious changes of mean exam score between the groups before and after the vetting process. Table 3 demonstrates the comparison of mean exam scores within each group in both stages (i.e. pre and post) according to item formats except SBA which did not fulfill assumptions of paired t-test. The mean exam scores of each group across the item formats were not significantly different between pre and post stages. In other words, the vetting process did not affect the mean exam score.

Table 2: Comparison of mean examination score between study groups (control and experimental) at each stage according to item formats.

Stage	Item type	Group (n)	Mean (SD)	Mean difference (95% CI)	t-statistics[a] (df)	p-value
Pre	MTF	Control (65) Experimental (64)	30.85 (5.23) 30.94 (4.61)	-0.09 (-1.81, 1.63)	-0.11 (127)	0.916
	SBA	Control (65) Experimental (64)	5.22 (1.87) 4.61 (1.72)	0.61 (-0.02, 1.23)	1.92 (127)	0.057
	EMQ	Control (65) Experimental (64)	5.09 (1.59) 5.19 (1.68)	-0.10 (-0.67, 0.47)	-0.33 (127)	0.741
	Overall	Control (65) Experimental (64)	41.15 (6.46) 40.73 (5.81)	0.42 (-1.72, 2.56)	0.39 (127)	0.699
Post	MTF	Control (65) Experimental (64)	31.57 (4.19) 30.64 (3.65)	0.93 (-0.44, 2.30)	1.34 (127)	0.182
	SBA	Control (65) Experimental (64)	5.51 (1.53) 4.75 (1.78)	0.76 (0.18, 1.34)	2.59 (127)	0.011
	EMQ	Control (65) Experimental (64)	5.42 (1.54) 5.03 (1.51)	0.38 (-0.15, 0.92)	1.43 (127)	0.155
	Overall	Control (65) Experimental (64)	42.49 (5.37) 40.42 (5.20)	2.07 (0.23, 3.91)	2.23 (127)	0.028

[a.] Independent t-test was applied

Table 3: Comparison of mean exam score within each group at both stages of examination (pre and post) according to item formats.

Group (n)	Item type	Stage	Mean (SD)	Mean difference (95% CI)	t-statistics[a] (df)	p-value
Control (65)	MTF	Pre Post	30.85 (5.23) 31.57 (4.19)	-0.72 (-2.13, 0.68)	-1.03 (64)	0.308
	EMQ	Pre Post	5.09 (1.59) 5.42 (1.54)	-0.32 (-0.76, 0.11)	-1.49 (64)	0.142
	Overall	Pre Post	41.15 (6.46) 42.49 (5.37)	-1.34 (-2.97, 0.30)	-1.64 (64)	0.107
Experimental (64)	MTF	Pre Post	30.94 (4.61) 30.64 (3.65)	0.30 (-0.85, 1.44)	0.52 (63)	0.606
	EMQ	Pre Post	5.19 (1.68) 5.03 (1.51)	0.16 (-0.32, 0.63)	0.65 (63)	0.516
	Overall	Pre Post	40.73 (5.81) 40.42 (5.20)	0.31 (-0.10, 1.60)	0.48 (63)	0.630

[a] Paired t-test was applied

Table 4 demonstrates the comparison of median exam score within each group at pre and post stages for SBA. The median examination score of each group was not significantly different between before and after the vetting process.

Table 4 Comparison of median exam score within each group at both stages of examination (pre and post) for SBA

Group (n)	Item type	Stage	Median (IQR)	Z-statistics[b]	p-value
Control (65)	**SBA**	**Pre**	5.00 (3)	-1.11	0.268
		Post	5.00 (2)		
Experimental (64)	**SBA**	**Pre**	4.00 (3)	-0.45	0.652
		Post	5.00 (3)		

[b] Wilcoxon Singed Ranked Test was applied

Table 5 demonstrates the comparison of pass-fail outcome between the intervention groups at each stage according to item formats. It showed no significant association between pass-fail outcome and the groups (control and experimental) except for SBA at the post stage (p-value = 0.008). The control group passing rate was better than experimental group and the failure rate of control group was lower than experimental group. Apart from this finding, the overall findings indicated no changes of pass-fail outcome between the groups before and after the vetting process

Table 5: Comparison of pass-fail outcomes between groups (control and experimental) at each stage according to item formats

Stage	Item type	Group (n)	Pass-Fail outcome (n) Fail	Pass	p-value
Pre	**MTF**	**Control (65)**	8	57	0.236 [a]
		Experimental (64)	4	60	
	SBA	**Control (65)**	28	37	0.187 [a]
		Experimental (64)	35	29	
	EMQ	**Control (65)**	24	41	0.624 [a]
		Experimental (64)	21	43	
	Overall	**Control (65)**	9	56	0.428 [a]
		Experimental (64)	6	58	
Post	**MTF**	**Control (65)**	4	61	1.000 [b]
		Experimental (64)	3	61	
	SBA	**Control (65)**	15	50	0.008 [a]
		Experimental (64)	29	35	
	EMQ	**Control (65)**	18	47	0.527 [a]
		Experimental (64)	21	43	
	Overall	**Control (65)**	4	61	0.136 [a]
		Experimental (64)	9	55	

[a] Pearson Chi-square test and [b] Fisher's exact tests were applied. Level of significance was set at 0.05

Table 6 shows the comparison of pass-fail outcome within each group at both stages according to item formats. There is no significant change of pass-fail outcome within each group at the pre and post stages except for control group for SBA (p-value = 0.019) where the passing rate of control group at post stage was higher than the pre stage. Overall, the vetting process did not change the pass-fail outcome of the experimental group between pre and post stages.

Table 6: Comparison of pass-fail outcomes within each group in both stages (pre and post) according to item formats.

Group (n)	Item type	Stage	Outcome	Post		p-value*
				Fail	Pass	
Control (65)	MTF	Pre	Fail	1	**7**	0.344
			Pass	**3**	54	
	SBA	Pre	Fail	8	**20**	0.019
			Pass	**7**	30	
	EMQ	Pre	Fail	9	**15**	0.307
			Pass	**9**	32	
	Overall	Pre	Fail	2	**7**	0.180
			Pass	**2**	54	
Experimental (64)	MTF	Pre	Fail	1	**3**	1.000
			Pass	**2**	58	
	SBA	Pre	Fail	21	**14**	0.286
			Pass	**8**	21	
	EMQ	Pre	Fail	9	**12**	1.000
			Pass	**12**	31	
	Overall	Pre	Fail	1	**5**	0.581
			Pass	**8**	50	

*Mc Nemar test was applied. Level of significance was set at 0.05

Discussion

The participants were randomly assigned into two equal groups through simple randomization technique. A similar method was utilised in a previous study done by Cizek (18). Maintaining homogeneous groups is an important issue in comparison (17). Results showed that the only difference between groups in this study was that the control group had more students with other entry qualification background compared to the experimental group. In addition, our data showed that the control group performed relatively better than the experimental group particularly in the examination performance. This finding could be due to mal-distribution of entry qualification between groups as a result of simple randomization method. This is in line with previous studies which reported that entry qualification predicted academic performance in medical school (19, 20). Nevertheless, other variables were successfully distributed randomly and equally between groups. From that notion, both groups were considered as homogenous. Future studies should apply stratified randomization technique to ensure equal distribution of entry qualification between groups.

Results (table 2, 3 and 4) demonstrated that the students' examination performance was not affected by the vetting process.. These findings

were dissimilar with previous studies which reported that flawed items (i.e. the non-vetted items) had negative impacts on students' examination performance (8, 12, 13). For example, Downing (8, 12) studied flawed and unflawed items. He found that flawed items tend to increase the failure rate of students more than the unflawed items. Tarrant and Ware (13) replicated Downing's study. They used ten tests. They found that students score higher with unflawed items than flawed items. However, it is worth highlighting that the previous studies were designed based on retrospective descriptive study, while our study utilized a randomized control trial (RCT) design which is a better and more robust design (17, 21). From that notion, our data should provide a stronger evidence of the impacts of the vetting process on students' performance during examination. Apart from that, interestingly we noted that, at the second mock examination (i.e. post stage), students in the control groups significantly obtained better marks than those in the experimental group in SBA as well as the overall scores. This was opposite to what was expected where the experimental group should perform better than the control group. A possible explanation for these unexpected results could be the heterogeneity of entry qualification background of students in the groups as had been discussed in the earlier section. Despite these unexpected findings, the vetting process failed to demonstrate any beneficial impact on students' performance in the examinations. Continued research in different education settings is required to verify the credibility of our findings.

Previous studies (8, 12, 13) reported contradictory findings regarding pass-fail outcomes. Downing (8, 12) found that flawed items (i.e. non-vetted items) were associated with lower passing rate and higher failure rate than unflawed items (i.e. vetted items), while Tarrant and Ware (13) found that unflawed items were associated with lower passing and higher failure rate than flawed items. In contrast, our data did not demonstrate any beneficial or harmful effects of the vetting process on the passing and failure rates in examinations either within the groups (i.e. changes) or between the groups (i.e. differences) (table 5 and 6). Notwithstanding the finding, we found that in specific occasions (i.e. SBA as discussed in previous section) control groups (i.e. those who received non-vetted items) had higher passing rate and lower failure rate than the experimental group (i.e. those who received vetted items). In a nutshell, these data demonstrated that the vetting process did not have any effect on the pass-fail outcome of students during examinations. We postulate that, this finding might hold true for medical schools that have an established faculty development program that trains their medical teachers on constructing good MCQ items (22, 23). Therefore, replicating this study on medical schools that do not have such faculty development program might be worthwhile. Nevertheless, continued research should done in different educational setup is required to verify our findings.

This study had several limitations that should be considered for future studies as well as for interpretation. The first limitation is related to the randomization technique in which this study applied a simple randomization method that may have lead to heterogeneous groups. It is recommended for future studies to apply a stratified randomization technique to ensure homogeneity of groups to control for certain confounding factors that might compromise results accuracy, for example the previous academic performance and entry qualification. The second limitation is related to the small sample of test items included in the study particularly for SBA and EMQ. Therefore, if this study is replicated in the future, calculation of appropriate sample size based on the present finding should be done. It is worth highlighting that this study was an initial effort to evaluate the current practice of test item review and hopefully will act as a baseline for future research. The third limitation is related to the confinement of this study to one batch of medical students. Therefore, it is better to conduct this study on different batches of students. Fourthly, this study was confined to a medical school to evaluate its own practice. To verify the present findings it is better to replicate this study in other schools either in the same university or other

universities. A replication of such a study in newly established medical schools with a limited experience in constructing and vetting test items may produce significant results.

The findings of this study brings the authors to a dilemma: on the one hand, in the face of decades of established thought, supported by numerous studies done by recognized experts in the field, to conclude that the vetting process is of no value is premature and bold, even arrogant. In this light, all efforts have been made to look at and discuss the possible weaknesses and limitations of the study. The authors are certainly aware that wisdom is certainly more than looking at numbers. But on the other hand, considering the efforts that were taken to ensure the robustness of the study, such as utilizing the best design for such objectives, etc, it cannot then simply be said that the results are worthless and should be chucked away. To do that would be to disrespect the tradition of scientific inquiry. Furthermore, the findings are not without support from other studies, and the 'established' findings are not immune to scrutiny and criticism. The way forward, then, would be to maintain a healthy, scientific, 'skeptical' attitude and look further into the issues that this study has highlighted; refine the components of vetting that do and do not impact on assessment quality and students' performance and derive practical applications of such findings. .

The authors still believe that the vetting process plays an important role to ensure high quality of test items constructed. Nevertheless, we venture to suggest that the vetting process can be limited to the departmental level rather than involving a central level. Formatting can be done by trained staff in the academic office using established guidelines to maintain uniformity of test items. Faculty training in test items construction as well as the vetting process should be maintained to ensure quality assessment in the medical school.

Acknowledgement

The authors wish to acknowledge the Islamic Development Bank Group and University of Science and Technology for sponsoring Dr. Majed Wadi throughout his study period, two years in study Master in Medical Education.

Funding

This study was funded by Universiti Sains Malaysia's Short Term Grant, with grant No 304/PPSP/61311089

Reference

1. Downing SM, Haladyna TM. Test Item Development: Validity Evidence From Quality Assurance Procedures. Applied Measurement in Education. 1997;10(1):61.
2. Haladyna TM, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Applied Measurement in Education. 2002;15(3):309 - 33.
3. Haladyna TM. Developing and validating multiple-choice test items. 3rd ed: Lawrence Erlbaum; 2004.
4. Downing SM. Twelve Steps for Effective Test Development. In: Downing SM, Haladyna TM, editors. Handbook of test development. 1st ed. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers; 2006. p. 3-25.
5. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a committee review process to improve the quality of course examinations. Advances In Health Sciences Education: Theory And Practice. 2006;11(1):61-8.
6. Tarrant M, Ware J. A framework for improving the quality of multiple-choice assessments. Nurse Educator. 2012;37(3):98.
7. Baranowski RA. Item Editing and Editorial Review. In: Downing SM, Haladyna TM, editors. Handbook of Test Development. 1st ed. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers; 2006. p. 349-57.
8. Downing SM. Construct-irrelevant Variance and Flawed Test Questions: Do Multiple-choice Item-writing Principles Make Any Difference? Academic Medicine. 2002;77(10)(Supplement):S103-S4.
9. Downing S. Threats to the Validity of Locally Developed Multiple-Choice Tests in Medical Education: Construct-Irrelevant Variance and Construct Underrepresentation. Advances in Health Sciences Education. 2002;7(3):235-41.

10. Downing SM. Validity: on the meaningful interpretation of assessment data. Medical Education. 2003;37(9):830-7.
11. Jozefowicz RFMD, Koeppen BMMDP, Case SP, Galbraith RMD, Swanson DP, Glew RHP. The Quality of In-house Medical School Examinations. Academic Medicine. 2002;77(2):156-61.
12. Downing SM. The Effects of Violating Standard Item Writing Principles on Tests and Students: The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education. Advances in Health Sciences Education. 2005;10(2):133-43.
13. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Medical Education. 2008;42(2):198-206.
14. IntelligenceSystemsSdnBhd. SmartScan. 5 ed. Kuala Lumpur, Malaysia2003.
15. SPSSInc. a. IBM SPSS Statistics 19. version 19.0.0 ed. Armonk, New York, USA2010.
16. StataCorpLP. Stata 9. 9.2 ed. College Station, Texas, USA2006.
17. Cohen L, Manion L, Morrison K, Morrison KRB. Research methods in education: Psychology Press; 2007.
18. Cizek GJ. The effect of altering the position of options in a multiple-choice examination. Educational and Psychological Measurement. 1994;54(1):8-20.
19. Hod R. Selection of Medical Students: the relationship between pre-admission academic achievements and students' profile to performance in medical school. Kubang Kerian, Malaysia: Universiti Sains Malaysia; 2006.
20. Yusoff MSB, Rahim AFA, Baba AA, Esa AR. Medical Student Selection Process and Its Pre-Admission Scores Association with the New Students' Academic Performance in Universiti Sains Malaysia. International Medical Journal. 2011;18(4):327-31.
21. Schuwirth L, Colliver J, Gruppen L, Kreiter C, Mennin S, Onishi H, et al. Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Medical Teacher. 2011;33(3):224-33.
22. Ja'afar R. Two decades of championing faculty development: Is it worth the effort? Education In Medicine Journal. 2012;4(2):1-6.
23. Ja'afar R. Monograph on Faculty Development at the School of Medical Sciences 3rd ed: Universitis Sains Malaysia; 2011.

Appendix I: Examples of MCQ before and after vetting

MCQ type	Before vetting	After vetting
MTF	Symptoms of metabolic alkalosis include A. Slow as well as shallow respiration B. Hyperactive reflexes C. Tetany D. Atrial tachycardia E. Dysrhythmias	Features of metabolic alkalosis include A. atrial tachycardia B. dysrhythmias C. hyperactive reflexes D. shallow respiration E. tetany
SBA	A 65-year-old man, a heavy smoker for 20 years presents with difficulty in swallowing solid food in the last one month. Oesophagoscopy reveals a polypoid mass projecting into the lumen of the middle third. A biopsy of the mass is taken. Which of the following is the most likely histopathological diagnosis ? A. Adenocarcinoma B. Oesophageal stricture C. Leiomyoma D. Squamous cell carcinoma E. Lymphoma	A 65-year old man complains of difficulty in swallowing solid food in the last one month. He has history of recurrent epigastric pain associated with retrosternal burning sensation. Oesophagoscopy reveals a polypoid mass projecting into the lumen of the lower third of oesophagus. A biopsy of the mass is taken. Which of the following is the most likely histopathological diagnosis? A. Adenocarcinoma B. Leiomyoma C. Lymphoma D. Oesophageal stricture E. Squamous cell carcinoma
EMQ	A. The glossopharyngeal nerve B. The recurrent laryngeal nerve C. The superior laryngeal laryngeal nerve D. The internal laryngeal nerve E. The external laryngeal nerve F. The inferior laryngeal nerve G. The pharyngeal branch of vagus nerve H. The pharyngeal branch of maxillary nerve I. The pharyngeal plexus 1. The nerve involves in sensory innervations of the nasopharynx 2. The nerve receives general sensation from the laryngeal cavity 3. The nerve that could be injured when a patient presented with hoarseness of voice after thyroid surgery	**Theme:** Innervations of the upper respiratory tract **A.** The external laryngeal nerve **B.** The glossopharyngeal nerve **C.** The inferior laryngeal nerve **D.** The internal laryngeal nerve **E.** The pharyngeal branch of maxillary nerve **F.** The pharyngeal branch of vagus nerve **G.** The pharyngeal plexus **H.** The recurrent laryngeal nerve **I.** The superior laryngeal laryngeal nerve. **For each statement below, select the most likely nerve involved** **1)** Mucosa lining around the opening of pharyngotympanic tube is innervated by this nerve **2)** A patient needing anaesthesia of the larynx from the epiglottis to the level of the vocal cords during bronchoscopy **3)** A patient presenting with hoarseness of voice following thyroid surgery