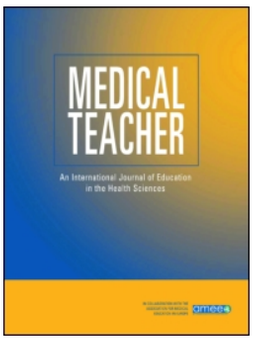

Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

Peer review improves psychometric characteristics of multiple choice questions

Hani Abozaid, Yoon Soo Park & Ara Tekian

To cite this article: Hani Abozaid, Yoon Soo Park & Ara Tekian (2017): Peer review improves psychometric characteristics of multiple choice questions, Medical Teacher, DOI: 10.1080/0142159X.2016.1254743

To link to this article: http://dx.doi.org/10.1080/0142159X.2016.1254743

Published online: 20 Jan 2017.

MEDICAL TEACHER, 2017
http://dx.doi.org/10.1080/0142159X.2016.1254743

MEDICAL TEACHER
Taylor & Francis
Taylor & Francis Group

Peer review improves psychometric characteristics of multiple choice questions

Hani Abozaid[a], Yoon Soo Park[b] and Ara Tekian[b]

[a]Department of Community Medicine, Faculty of Medicine, Taif University, Saudi Arabia; [b]Department of Medical Education, University of Illinois College of Medicine at Chicago, Chicago, IL, USA

ABSTRACT

Purpose: For new and emerging medical schools, developing a system to peer-review and evaluate the assessment processes through faculty development programs can be a challenge. This study evaluates the impact of peer-review practices on item analysis, reliability, and the standard error of measurement of multiple-choice questions for summative final examinations.

Methods: This study used a retrospective cohort design of two consecutive academic years in 2012 and in 2013. Psychometric analyses of multiple-choice questions of three summative final examinations in Medicine, Pediatrics, and Surgery for sixth year medical students at the College of Medicine Taif University were used. Formal peer review of multiple-choice questions began in 2013, using guidelines from the National Board of Medical Examiners. Psychometric analyses of multiple-choice questions included item analysis (item difficulty and item discrimination) and calculation of internal-consistency reliability and the standard error of measurement. Data analyses were conducted using Stata.

Results: Results showed significant improvement in psychometric indices, particularly item discrimination and reliability by .14 and .12 points, respectively, following the implementation of the peer review process across the three exams. Item difficulty remained unchanged for Pediatrics and Surgery.

Conclusion: Peer-review practices of multiple-choice questions using guidelines can lead to improved psychometric characteristics of items; these findings have implications for faculty development programs in improving item quality, particularly for medical schools in early stages of transforming assessment practices.

Introduction

Assessment of the medical student performance is an important part of the educational framework in medical colleges. Assessment scores can be used for motivation, promotion, and feedback for areas of weakness in medical students (American Educational Research Association et al. 2014). Generally, assessments can be designed to measure three different levels of learning outcomes: knowledge, skills and attitudes. In this respect, various assessment methods such as multiple choice questions (MCQs), written essays, standardized patient encounters, and direct observation tools can be used.

MCQs are an important tool for formative and summative assessment that can measure and assess the knowledge of learners in medical schools. MCQs can be designed to measure not only knowledge, but also higher-order diagnostic reasoning, as well as application, integration, and synthesis (Vanderbilt et al. 2013). However, medical schools must review the quality of MCQs to ensure their validity and reliability. In particular, MCQs that violate principles of evidence-based effective item writing guidelines – unfocused stem, negatives, or options that include "all of the above" or "none of the above" – contribute to construct irrelevant variance, thereby threatening the validity of the assessment (Case & Swanson 2002; Haladyna 2004). Such flawed items fail to provide validity evidence in the psychometric characteristics of the assessment; moreover, flawed items have been shown to affect pass rates for some students (Downing 2005). As such, it is essential to follow evidence-based guidelines for writing valid and reliable MCQs to maintain standards for quality (Haladyna et al. 2002; American Educational Research Association et al. 2014).

Standards for item writing have called for training test developers who can improve the quality of items (Jozefowicz et al. 2002). Within this context, faculty development programs (FDPs) can be designed to support faculty members in MCQ writing and evaluation of item quality. Studies in the literature have supported such FDPs, and they include evidence showing their impact. A prior study conducted at Riphah University in Pakistan showed

Practice points

- Implementing peer-review systems for multiple choice questions can significantly improve the overall psychometric quality of local summative examinations.
- Our findings show that following best practice guidelines, such as the one offered by the National Board of Medical Examiners, can serve to improve test quality.
- Results of this study indicate that peer-review systems can be effective even for young and emerging medical schools.

CONTACT Hani Abozaid, MD drhaniabozaid@yahoo.com Department of Community Medicine, Faculty of Medicine, Taif University, Taif, Al-Haweiah, P.O. Box 888, 21974, Saudi Arabia

that item flaws affected the quality of MCQs; however, identifying flaws in MCQs by frequent feedback from faculty and assigned committees improved the quality and ultimately decreased the number of item flaws (Humaira et al. 2013). Another study conducted in nursing at the University of Hong Kong showed that nearly half of all items administered in 10 exam forms had some form of item flaw; the authors recommended that providing FDPs on proper item writing, and initiating item reviews before and after the exam could improve the quality of high stake exams (Tarrant & Ware 2008). In addition, studies conducted in Australia and in the United States have also shown that implementing policies for MCQ review can improve the psychometric quality of the exams (Wallach et al. 2006; Malau-Aduli & Zimitat 2011).

While these studies demonstrate the value of FDPs and the effect of implementing psychometric review processes as means to improve the quality of MCQs, it is unclear what impact they can have particularly for new and emerging medical schools – the challenges associated with emerging medical schools differ from institutions with longer history of FDPs or access to resources. An approach to deliver FDPs is to conduct peer review sessions that consist of three parts – (1) training faculty to become peer reviewers, (2) training of faculty to construct MCQs of quality, and (3) enhancing the faculty skills through feedback they receive from the peer reviews. These peer reviews require faculty to be trained on standards of MCQ writing (Case & Swanson 2002) and interpreting post-exam item analysis, including the understanding of item characteristics (item difficulty, item discrimination), reliability indices, and performance of distractor options. These methods can help identify effective items that discriminate performance among students and eliminate nonfunctioning distractors in ensuing examinations (Tarrant et al. 2009). Item analysis informs the quality of items to evaluate whether student performance aligns with the intended utility of the items hypothesized by test developers (Shakil 2008). Without item analysis, it is challenging to identify which items are effective and which items are poor performing (i.e. does not provide information on students' knowledge level and does not discriminate differences between high and low performing learners).

This study evaluates the effect of implementing peer review practices of item analysis, reliability, and standard error of measurement (SEM) of MCQs in the final summative examinations.

Methods

Organizational context of medical collage at Taif University

The College of Medicine (COM) at Taif University was established in 2005 and is one of the newest medical schools in Saudi Arabia. The COM follows a six-year undergraduate, integrated curriculum. It has three Phases. Phase One (first year) students study general sciences; Phase Two (second year, third year, and first semester of the fourth year) students go through a systems-based curriculum; and Phase Three (second semester of the fourth year, fifth year, and sixth year) students start clinical rotations. Assessments in Phase Three are based on a blueprint, using different types of assessment methods, including MCQs, objective structured clinical examination (OSCE), and other clinical examinations.

Faculty development through peer review

The initial faculty development in assessment began as part of an international collaboration between Taif Medical College and the University of Illinois at Chicago (UIC) in 2012. Faculty from UIC conducted workshops on test development, item writing, and item analysis for all academic staff. This workshop aimed to improve the quality of item writing and decrease the number of item flaws, following evidence-based international guidelines.

In 2013, these workshops led to the development of a policy by the Medical Education Department at Taif University to conduct peer reviews of MCQs and to provide feedback to item writers and course faculty. The main objectives of this policy were to improve the psychometric quality of the high-stakes exams and to decrease the number of item flaws. This policy stated that the Medical Education Department would assign a committee for each exam to review MCQs and to provide feedback to the course faculty for improvement. The feedback provided by the peer-review committee members to the course faculty functioned as faculty development sessions that (1) improved the exam quality and (2) enriched the course faculty's understanding of assessment and MCQ writing principles.

The MCQ peer-review committee reviewed the newly-developed MCQs from each department, which was sent one week before administering the exam. The MCQ peer-review committee consisted of seven members: (1) Three permanent members from the Exam Unit, (2) one member from the Medical Education Department, (3) one member from the Scientific Committee, and (4) two faculty members from the assigned department. The member from Medical Education Department worked as facilitator and provided guidance for the committee.

Based on National Board of Medical Examiners (NBME) guidelines (Case & Swanson 2002), this committee identified item flaws, recommended removal or medication of items using the guideline, and provided feedback and recommendations to the item writer and faculty for future improvement of item writing. The Medical Education Department and Scientific Committee are essential to the blueprinting and organization of the assessment in Phase One, Phase Two, and Phase Three.

Data

Study design

This study used a retrospective cohort design of two consecutive academic years in 2012 and in 2013. Data from 2012 were used as baseline reference to compare with data from 2013, following the implementation of the peer-review FDP.

Data collection

Psychometric analyses of MCQs of three summative final examinations in Medicine, Pediatrics, and Surgery for sixth year (Phase Three) medical students at COM Taif University were used. These three courses are essential for sixth-year

medical students and have higher credit hours than other courses; summative exams in these courses are considered high stakes, which requires maintenance of psychometric quality.

MCQs ranged from 52 to 110 single selected-response items, with four or five options. The MCQ peer-review committee reviewed all items to check for flaws that violate the NBME guidelines. In addition, psychometric analyses of MCQs for item difficulty (proportion correct) and item discrimination (point-biserial correlation) were calculated as part of item analysis. Furthermore, internal-consistency reliability (Cronbach's alpha) and the standard error of measurement (SEM) were calculated. All psychometric analyses were conducted using Stata 14 (StataCorp, College Station, Texas).

Peer-review: item analysis

Item analysis was based on review of item difficulty and item discrimination. In addition, using these indices, reliability and SEM were reviewed as part of the peer-review process. The item difficulty index measures the ratio of learners who selected the correct answer and all test takers. It ranges from 0 to 1, where 1 indicates all students answered the item correctly; 0 indicates that no student got the item correct. As such, the lower the item difficulty, the item is more difficult (Shakil 2008).

Item discrimination measures how well the item is able to discriminate differences between students of high and low ability. It ranges from −1 to +1, wherein 0 means no discrimination, +1 indicates good discrimination and −1 means those who achieved low scores answered better than those who achieved high score, which can be a threat to the validity of items. According to Downing and Yudkowsky (2009), an acceptable range for item discrimination is greater than 0.20.

MCQ examinations typically use measures of internal-consistency reliability to examine the reproducibility of assessment scores. Cronbach's alpha is used for this purpose, which ranges between 0 and 1. Low reliability may indicate inconsistent decisions of the learner depending on the MCQs used in the assessment. Generally, for MCQs, reliability above .70 is expected (Nunnally 1978; Park et al., 2016).

The SEM is a function of the reliability and the variability of the test scores and can be used to determine the measurement precision of the assessment. Assessments with larger SEM indicate lower precision, while assessments with lower SEM indicate higher precision (Park et al., 2016).

Analysis

Data compilation and analyses were conducted using Stata 14 (StataCorp, College Station, Texas). Independent *t*-tests were used to compare differences between means; *p*-value of less than .05 was used to determine significance. Institutional Review Board (IRB) at Taif University approved this study.

Results

Descriptive statistics

Table 1 shows the descriptive statistics for the three summative examinations (Medicine, Pediatrics, and Surgery) taken by sixth-year medical students in 2012 and in 2013.

Table 1. Descriptive statistics: course scores by year.

Course	Score	Year 2012		Year 2013	
Medicine	Mean Raw Score	72.16	(8.31)	61.38	(18.26)
(2012: $n = 103$)	Mean % Score	72.16	(8.31)	61.38	(18.26)
(2013: $n = 102$)	% Pass	92.23		63.73	
Pediatrics	Mean Raw Score	31.50	(5.73)	64.50	(16.50)
(2012: $n = 102$)	Mean % Score	60.58	(11.01)	64.50	(16.50)
(2013: $n = 94$)	% Pass	50.98		61.70	
Surgery	Mean Raw Score	62.03	(12.76)	54.30	(16.64)
(2012: $n = 106$)	Mean % Score	56.39	(11.60)	54.30	(16.64)
(2013: $n = 109$)	% Pass	47.17		35.78	

Values in parenthesis are standard deviations; "*n*" refers to the number of examinees.

Medicine

In 2012, 103 medical students took the Medicine final examination, where 92.23% of them pass the exam (a 60% institutional cut score was applied). The mean score was 72.16%, and Standard Deviation (SD) was 8.31% (Min = 47%, Max = 87%). In 2013, following peer review, 102 medical students took the Medicine final examination, where 63.73% passed the exam (applying the same 60% institutional cut score). The mean score was 61.38% (SD = 18.26%, Min = 17%, Max = 91%).

Pediatrics

In 2012, 102 medical students took the pediatrics final examination, where 50.98% passed the exam (Mean = 60.58% SD = 11.01%, Min = 32.69, Max = 82.69%). In 2013, 94 took the exam, where 61.70% passed (Mean = 64.50%, SD = 16.50%, Min = 21%, Max = 92%).

Surgery

In 2012, among 106 medical students who took the surgery examination, 47.17% passed (Mean = 56.39, SD = 11.60%, Min = 20.91%, Max = 77.27). In 2013, 109 medical students took the exam, where 35.78% passed (Mean = 54.30%, SD = 16.64%, Min = 15%, Max = 87%).

Psychometric characteristics of items

In 2012, out of 100 items from the Medicine examination, the mean item difficulty was .72, while the mean item discrimination was .22. Reliability was .84, and SEM was 3.28. In 2013, out of 100 items, mean item difficulty was .61, and mean item discrimination was .39, indicating significant change in item difficulty, $p = .002$, and improvement in item discrimination, $p < .001$. Moreover, reliability increased to .95, while SEM shifted to 4.10. Table 2 summarizes these results for medicine, pediatrics, and surgery.

For pediatrics, item difficulty changed from .61 to .65 between 2012 and 2013 (not significant). However, item discrimination improved significantly from .20 to .36, $p < .001$. This resulted in improved reliability from .74 to .94. SEM increased from 2.90 to 4.02 due to nearly doubling test length.

Finally for surgery, there was no significant change in item difficulty between 2012 and 2013; however, item discrimination improved from .24 to .34. This also increased reliability from .88 to .93. SEM decreased from 4.41 to 4.32.

Overall, item discrimination for the three assessments improved significantly, resulting in higher reliability indices.

Table 2. Item analysis, reliability, and standard error of measurement (SEM) by course and year.

Course	Statistic	Year 2012		Year 2013		*p*-value
Medicine	Item Difficulty	.72	(.29)	.61	(.19)	.002
(2012: items = 100)	Item Discrimination	.22	(.17)	.39	(.15)	<.001
(2013: items = 100)	Reliability	.84		.95		
	SEM	3.28		4.10		
Pediatrics	Item Difficulty	.61	(.27)	.65	(.19)	.290
(2012: items = 52)	Item Discrimination	.20	(.19)	.36	(.16)	<.001
(2013: items = 100)	Reliability	.74		.94		
	SEM	2.90		4.02		
Surgery	Item Difficulty	.56	(.24)	.54	(.19)	.507
(2012: items = 110)	Item Discrimination	.24	(.18)	.34	(.14)	<.001
(2013: items = 100)	Reliability	.88		.93		
	SEM	4.41		4.32		

Values in parenthesis are standard deviations.

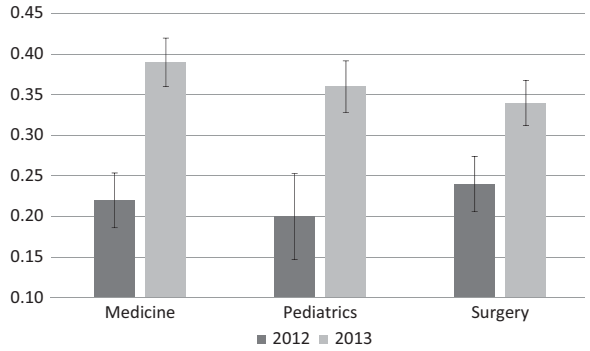

Figure 1. Item discrimination of test items by course and year: mean item discrimination ±95% confidence interval. Note: Error bars indicate 95% confidence intervals.

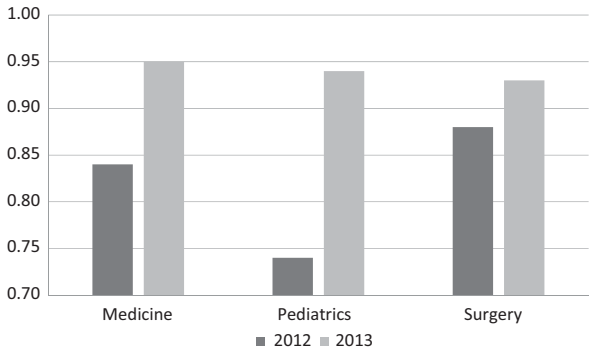

Figure 2. Reliability (Cronbach's alpha) of test items by course and year.

Figures 1 and 2 were added to illustrate these changes in psychometric indices.

Discussion

This study presents empirical evidence that the peer-review process of MCQs can significantly improve the psychometric quality of items. These results can be particularly meaningful given the context that Taif University Medical College is a relatively young medical school undergoing changes to its curricular and assessment practices. And as such, the peer-review process provided an essential quality assurance measure. Overall, findings from this study, based on three main assessments from Medicine, Pediatrics, and Surgery, across two academic years in 2012 (no peer review) and 2013 (peer review implemented) showed evidence of significant improvement in item discrimination, which in turn improved reliability. For Pediatrics and for Surgery, the improvement in psychometric quality was made without significantly altering item difficulty. In other words, the assessments were able to better discriminate between test takers and yield more reproducible test scores, while maintaining the overall test difficulty. Item discrimination and reliability have a direct relationship, where increase in discrimination improves reliability, and decrease in item discrimination may reduce reliability.

Results from this study reiterate findings from prior studies conducted at the School of Medicine, University of Tasmania in Australia, to measure the effect of peer review practices of multiple-choice examinations in the first three years. The researchers found that peer-review process led to a decrease in items with low discrimination; at the same time, peer-review process led to higher item discrimination and higher reliability (Malau-Aduli & Zimitat 2011). Another study conducted at the University of South Florida College of Medicine concluded that guidelines and review processes improved item characteristics for local in-house assessments (Wallach et al. 2006).

The findings in this current study were meaningful in that we examined the effect of peer-review process for three summative course exams across two full years. Only after a year of implementing the peer-review process, we were able to show significant improvements in our test quality, which would inform the overall validity of our test scores. The basis for the peer-review process was the use of NBME guidelines, which provide best practices for psychometric review and test development. Simply implementing these guidelines led to these results, perhaps indicating the robust impact that best practices can offer. These findings are also meaningful in that these results come from one of the newest medical schools in Saudi Arabia. This study and previous literatures provide strong recommendation to medical schools in general and young medical schools, in particular, to conduct peer review of items and follow NBME guidelines, for improving the quality of the items and having a valid and reliable assessment tool.

Limitations of the study

Findings from this study are based on data from a single institution with limited sample size. Moreover, the study used a retrospective cohort design, comparing data from 2012 as baseline with data from 2013 as intervention, in which students may not be completely randomized. In addition, this study only relies on internal reviewers and no external reviewers or benchmark with other national medical collages was used, which could limit the generalizability of the study. However, we used data from three of the largest summative examinations, providing evidence from multiple sources. Moreover, implications of quality assurance and peer reviewing can be particularly meaningful for other new and emerging schools. Future studies may use data from larger medical student settings to draw more generalizable conclusions. Efforts are underway to continue examining the longitudinal impact of peer review processes and to identify other potential barriers and challenges to continue improving our assessment system.

Conclusion

The peer review practices of MCQs based on NBME guidelines can lead to significant improvements in the quality of

items that can ultimately improve the validity of assessment scores. These findings reinforce known practices in assessment that can be applied even for new and emerging medical schools in the Gulf region.

Disclosure statement

The authors reports no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Glossary

Psychometric analysis: The analysis of psychological tests and measurements to ensure that scores are as reliable and valid as possible.

Notes on contributors

Dr Hani Abozaid, MD, is Associate Professor in the Department of Community Medicine, Faculty of Medicine, Taif University, Taif, Saudi Arabia.

Dr Yoon Soo Park, PhD, is Assistant Professor in the Department of Medical Education, College of Medicine, University of Illinois, Chicago, Illinois, USA.

Dr Ara Tekian, PhD, MHPE, is Associate Dean for International Affairs and Professor in the Department of Medical Education, College of Medicine, University of Illinois, Chicago, Illinois, USA.

Funding

The publication of this supplement has been made possible with the generous financial support of the Dr Hamza Alkholi Chair for Developing Medical Education in KSA.

Ethical approval

This institutional review board approved this study.

ORCID

Yoon Soo Park http://orcid.org/0000-0001-8583-4335

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. 2014. Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Case SM, Swanson DB. 2002. Constructing written test questions for the basic and clinical sciences 2nd ed. Philadelphia, PA: National Board of Medical Examiners.

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

Downing SM. 2005. 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. Adv Health Sci Educ Theory Pract. 19:133–143.

Haladyna TM, Downing SM, Rodriguez MC. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ. 15:309–333.

Haladyna TM. 2004. Developing and validating multiple-choice test items. Hillsdale, NJ: Lawrence Erlbaum Associates.

Humaira K, Khalid D, Azra A, Masood A. 2013. Identification of technical item flaws leads to improvement of the quality of single best Multiple Choice Questions. Pak J Med Sci. 29:715–719.

Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. 2002. The quality of in-house medical school examinations. Acad Med. 77:156–161.

Malau-Aduli BS, Zimitat C. 2011. Peer review improves the quality of MCQ examinations. Assess Eval High Educ. 37:1–13.

Nunnally JC. 1978. Psychometric theory. 2nd ed. New York: McGraw-Hill.

Park YS, Hyderi A, Bordage G, Xing K, Yudkowsky R. 2016. Inter-rater reliability and generalizability of patient note scores using a scoring rubric based on the USMLE Step-2 CS format. Adv in Health Sci Educ. 21:761–773.

Shakil M. 2008. Assessing student performance using test item analysis and its relevance to the state exit final exams of MAT0024 Classes: an action research project. Polygons. 2:1–35.

Tarrant M, Ware J, Mohammed AM. 2009. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ. 9:40

Tarrant M, Ware J. 2008. Assessment impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ. 42:198–206.

Vanderbilt A, Feldman M, Wood IK. 2013. Medical education: a review of course exams. Med Educ Online. 1:1–5.

Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. 2006. Use of a committee review process to improve the quality of course examinations. Adv Heal Sci Educ. 11:61–68.