Evaluation of an Intervention to Improve Quality of Single-best Answer Multiple-choice Questions

Kevin R. Scott, MD, MSEd* *Perelman School of Medicine at the University of Pennsylvania, Department of Emergency Medicine, Philadelphia, Pennsylvania
Andrew M. King, MD† †The Ohio State University Wexner Medical Center, Department of Emergency Medicine, Columbus, Ohio
Molly K. Estes, MD‡ ‡Loma Linda University Medical Center, Department of Emergency Medicine, Loma Linda, California
Lauren W. Conlon, MD*
Jonathan S. Jones, MD§ §Merit Health Central, Department of Emergency Medicine, Jackson, Mississippi
Andrew W. Phillips, MD, MEd¶ ¶University of North Carolina, Department of Emergency Medicine, Chapel Hill, North Carolina

Section Editor: Jonathan Fisher, MD, MPH
Submission history: Submitted July 13, 2018; Revision received November 4, 2018; Accepted November 8, 2018
Electronically published December 3, 2018
Full text available through open access at http://escholarship.org/uc/uciem_westjem
DOI: 10.5811/westjem.2018.11.39805

Introduction: Despite the ubiquity of single-best answer multiple-choice questions (MCQ) in assessments throughout medical education, question writers often receive little to no formal training, potentially decreasing the validity of assessments. While lengthy training opportunities in item writing exist, the availability of brief interventions is limited.

Methods: We developed and performed an initial validation of an item-quality assessment tool and measured the impact of a brief educational intervention on the quality of single-best answer MCQs.

Results: The item-quality assessment tool demonstrated moderate internal structure evidence when applied to the 20 practice questions (κ=.671, p<.001) and excellent internal structure when applied to the true dataset (κ=0.904, p<.001). Quality scale scores for pre-intervention questions ranged from 2-6 with a mean ± standard deviation (SD) of 3.79 ± 1.23, while post-intervention scores ranged from 4-6 with a mean ± SD of 5.42 ± 0.69. The post-intervention scores were significantly higher than the pre-intervention scores, $x^2(1) = 38$, p <0.001.

Conclusion: Our study demonstrated short-term improvement in single-best answer MCQ writing quality after a brief, open-access lecture, as measured by a simple, novel, grading rubric with reasonable validity evidence. [West J Emerg Med. 2019;20(1)11-14.]

INTRODUCTION

The use of single-best answer multiple-choice questions (MCQ) in examinations is ubiquitous in medical education. Although guidelines for writing MCQs exist, item writers often receive little to no formal training, potentially reducing the validity of examinations by introducing construct-irrelevant variance.[1-3] Extended educational interventions in the area of item writing have been shown to improve written item quality with shorter interventions showing a similar impact.[4-6] The literature suggests learners involved in item writing find it to be a positive learning experience that potentially improves performance on a summative assessment.[7-10]

The National Board of Medical Examiners (NBME) provides both a detailed, open-access guide for exam-question writing and an online training module.[11-13] These tools provide instruction for writing high quality MCQs and are used in the design of basic and clinical science exams, but

they are lengthy and oriented toward experienced question writers. Other tools remain lengthy and either require in-person workshops or are designed for self-study and require a prerequisite of basic question-writing understanding. Additionally, the literature lacks a simple MCQ quality metric with strong validity evidence. The two objectives of this study were to 1) establish validity evidence for a novel MCQ evaluation tool, and 2) evaluate the efficacy of a brief didactic lecture on MCQ question writing.

METHODS

Study Setting and Participants

We sought student and resident volunteers from the American Academy of Emergency Medicine Resident and Student Association, and conducted the educational intervention in September 2017. The study was granted exemption status by the University of Pennsylvania Institutional Review Board.

Multiple-choice Question Quality Assessment Tool Derivation

We created a MCQ quality assessment tool based on expert opinions (AWP, KRS, JJ) of the most important components contained in the question-writing lecture; it is based on multiple, well-accepted sources, supporting content evidence.[11,12] Two of the experts have formal education backgrounds including master's degrees (AWP and KRS) that included advanced training in item writing and quality assessment. The third expert (JJ) has taught question writing for several years to national audiences. We followed current standards that endorse validity based on Messick's model.[14-16] We created six items, each rated on a binary "present" or "not present" scoring system with a total minimum potential scale score of zero and a maximum potential scale score of six (Figure). Two additional educators (AK and ME) reviewed the rubric and shared their interpretations, which were aligned with the item objectives, supporting response-process evidence. A set of 20 questions with intentional errors was created (AWP), available in Appendix A, for the initial validity evidence assessment.

Training Module Creation and Assessment of Impact

The training module was created by an item-writing expert (JJ) using PowerPoint (Microsoft Corporation, Redmond, WA) with recorded voice-over (iMovie, Apple Inc., Cupertino, CA), allowing for independent completion by learners. The training module itself has been previously published in an open- access curriculum database and was based on principles of item writing as described by the NBME.[4,11,12,17] Participants were asked to write three novel, single-best answer MCQs based on a two-page excerpt from an emergency medicine board review textbook about trauma just prior to the lecture. They then watched the question-writing lecture together on YouTube (Google Inc., Mountain View, CA) on a conference call followed by a 10-minute question and answer period with a question-writing expert different than the lecturer (AWP). Participants were then asked to write three new, single-best answer MCQs based on the same excerpt immediately after the lecture.

Pre- and post-intervention MCQ quality scores were determined by two item-writing experts (AK, ME) via the item quality assessment tool. Discrepancies were decided by a third item-writing expert (LC).

Positively worded stem (0=no, 1=yes)

Stem phrased as a question (0=no, 1=yes)

Five answer choices (0=no, 1=yes)

Answer choices are listed alphabetically (0=no, 1=yes)

Foils are similarly complex as answer (0=no, 1=yes)

One clear, correct answer (0=no, 1=yes)

Figure. Multiple-choice questions quality assessment tool.

Statistical Analysis

We first performed descriptive summaries including mean and standard deviation (SD), frequencies, and total responses. Internal reliability was assessed using Cohen's kappa. We decided a priori to compare pre- and post-lecture scores using the non-parametric Friedman's analysis of variance (ANOVA), given the expected range to be relatively small and low likelihood of having an even distribution of the standard error of the mean. Friedman's ANOVA is essentially a non-parametric, repeated measures one-way ANOVA. A p-value less than 0.05 was considered statistically significant. We performed all analyses using SPSS version 24 (IBM Corporation, Armonk, NY).

RESULTS

Multiple-choice Question Quality Assessment Tool Validity Evidence

The internal structure evidence was moderate when the tool was applied to the 20 practice questions (κ=.671, p<.001). The tool demonstrated excellent internal structure when applied to the true dataset of questions created by the students and residents (κ=0.904, p<.001) with only eight discrepancies in 264 cases (48 total requested questions – 4 missing questions = 44 total questions with 6 points each yielding 264 cases), evaluated by two different researchers. Evidence of consequence was demonstrated as part of the other primary objective of this study, in which pre- and post-lecture scores were different. As this was a stand-alone study, we were unable to evaluate for relationships with other variables.

Training Module Impact on Item Quality

A total of eight residents and students consented and participated in the lesson, of whom seven provided both pre- and post-lecture MCQs. One participant provided two pre-

lecture questions rather than three, and another provided no post-lecture questions, thus totaling four total missing questions of the 48 possible total questions (8 x 3 x 2). Missing questions were excluded pairwise since the post questions were edits of the original questions; therefore, four missing questions led to elimination of eight total questions. We analyzed a total of 40 questions (20 pre- and 20 post-lecture). The MCQ quality scale scores for pre-intervention questions written by the learners ranged from 2 - 6 with a mean ± SD of 3.79 ± 1.23, while post-intervention scores ranged from 4 – 6. The post-intervention scores were significantly higher than the pre-intervention scores, $x^2(1) = 38$, $p < 0.001$.

DISCUSSION

The current study supports the efficacy of a short, high-yield lecture to teach best evidence in developing single-best answer MCQs. The study also provides strong validity evidence for a novel tool by which to evaluate the structure of single-best answer MCQs.

Although multiple prior studies have evaluated outcomes from an educational intervention to improve MCQ writing, the current study is the first available remotely, free to the public, and at approximately 30 minutes in length is the shortest.[5,6,18] These differences are important because this efficacious education intervention is replicable in any setting, whereas in-person workshops may vary with the instructor, size of the group, and other factors. The open-access availability through the educational platform at the *Journal of Education and Teaching in Emergency Medicine* (JETem) and its brief duration provide a practical advantage to this educational intervention as well.[17] Future work should directly compare other tools against this one.

Another important contrast to prior studies is the target group. Much focus has been placed on faculty development, yet educators are seeing the benefits of learners writing questions.[5-9,18-21] To this end, the current educational intervention was specifically designed for novice MCQ writers and tested in a sample of students and residents. It can be easily adopted by clerkship directors and program directors to use with students and residents as both a learning tool and as preparation to write questions as junior faculty members in future years.

This study lastly provides a checklist with reasonable validity evidence and strong inter-rater reliability when applied to the real-world questions. This is in contrast to other checklists that exist but are limited to content validity by experts.[18] It is unclear why the instrument had better inter-rater reliability with the real questions than when applied to the sham questions. We suspect this finding simply uncovered the inherent limitation of sham tests in which the author was trying to elicit specific flags in the tool. The strong performance with the live questions is reassuring.

LIMITATIONS

Our study must be interpreted in the context of several limitations. Most importantly, we studied a short-term outcome. This variable must be a precursor to follow-up, long-term learning outcomes to fully elucidate the efficacy of the intervention. It is also important to highlight that the intervention and assessment tool are intended to improve the structure of MCQs. Such proper practices are associated with good question quality as ascertained through psychometric analysis, but they are beyond the scope of our initial study. Additionally, our study recruited volunteers who may have been more motivated to improve their MCQ writing skills than students and residents in the general population. Finally, although the MCQ quality tool was applied against a test group of questions and a real-life group of questions, it was nonetheless a small sample of questions with a small number of participants, and the tool should be tested against more questions and more raters.

CONCLUSION

Our study demonstrated short-term improvement in single-best answer MCQ writing quality after a brief, open-access lecture, as measured by a simple, novel, grading rubric with reasonable validity evidence.

ACKNOWLEDGEMENT

We would like to acknowledge the American Academy of Emergency Medicine (AAEM) and the AAEM Resident and Student Association (AAEM/RSA) for coordinating the study participants and education session.

Address for Correspondence: Kevin R. Scott, MD, MSEd, Perelman School of Medicine at the University of Pennsylvania, Department of Emergency Medicine, 3400 Spruce Street, Ground Ravdin Philadelphia, PA 19104. Email: kevin.scott@uphs.upenn.edu.

Conflicts of Interest: By the *West*JEM article submission agreement, all authors are required to disclose all affiliations, funding sources and financial or management relationships that could be perceived as potential sources of bias. Dr. Phillips reports one disclosure as editor in chief at EM Coach LLC. Items written using this particular educational intervention may be incorporated into this board-review product.

REFERENCES

1. 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. *Adv Health Sci Educ*.

2005;10(2):133-43.
2. Ali SH, Ruit KG. The Impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. *Perspect Med Educ*. 2015;4(5):244-51.
3. Rush BR, Rankin DC, White BJ. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Med Educ*. 2016;16(1):250.
4. Webb EM, Phuong JS, Naeger DM. Does educator training or experience affect the quality of multiple-choice questions? *Acad Radiol*. 2015;22:1317-22.
5. Abdulghani HM, Ahmad F, Irshad M, et al. Faculty development programs improve the quality of multiple choice Questions items' writing. *Sci Rep*. 2015;5(1):9556.
6. Dellinges MA, Curtis DA. Will a short training session improve multiple-choice item-writing quality by dental school faculty? A pilot study. *J Dent Educ*. 2017;81(8):948-55.
7. Chamberlain S, Freeman A, Oldham J, et al. Innovative learning: employing medical students to write formative assessments. *Med Teach*. 2006;28(7):656-9.
8. Nwosu A, Mason S, Roberts A, et al. The evaluation of a peer-led question-writing task. *Clin Teach*. 2013;10(3):151-4.
9. Harris BHL, Walsh JL, Tayyaba S, et al. A novel student-led approach to multiple-choice question generation and online database creation, with targeted clinician input. *Teach Learn Med*. 2015;27(2):182-8.
10. Walsh J, Harris B, Tayyaba S, et al. Student-written single-best answer questions predict performance in finals. *Clin Teach*. 2016;13(5):352-6.
11. Case SM, Swanson DB. (2002). *Constructing Written Test Questions For the Basic and Clinical Sciences*. (3rd ed.). Philadelphia, PA: National Board of Medical Examiners.
12. Raymond M, Roeder C. National Board of Medical Examiners: Writing Multiple Choice Questions: An Introductory Tutorial. 2012. Available at: http://download.usmle.org/IWTutorial/MCQs_Intro_Tutorial.pdf. Accessed September 5, 2017.
13. Raymond M, Roeder C. National Board of Medical Examiners: Writing Multiple Choice Questions: An Introductory Tutorial. Available at: http://download.usmle.org/IWTutorial/intro.htm. Accessed September 5, 2017.
14. American Educational Research Association; American Psychological Association; National Council on Measurement in Education Association AER. Validity. (2014). *The Standards for Educational and Psychological Testing*. (11-31). Washington, DC: Amer Educational Research Assn.
15. Messick S. Validity. Linn RL, ed. (1989). *Educational Measurement*. (3rd ed.). (13-103). New York, NY: The American Council on Education/Macmillan.
16. Cook DA, Kuper A, Hatala R, et al. When assessment data are words. *Acad Med*. 2016;91(10):1359-69.
17. Jones JS, Phillips AW, King AM, et al. A brief didactic intervention to improve multiple-choice item-writing quality. *J Educ Teach Emerg Med*. 2018;3(1):L1-16. Available at: http://jetem.org/multiple_choice_didactic/. Accessed September 5, 2017.
18. Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. *Adv Heal Educ Theory Pract*. 2012;17(3):369-76.
19. Kim J, Chi Y, Huensch A, et al. A case study on an item writing process: use of test specifications, nature of group dynamics, and individual item writers' characteristics. *Lang Assess Q*. 2010;7(2):160-74.
20. Tunks J. The effect of training in test item writing on test performance of junior high students. *Educ Stud*. 2001;27(2):129-42.
21. Walsh JL, Denny P, Smith PE. Encouraging maximal learning with minimal effort using PeerWise. *Med Educ*. 2015;49(5):521-2.