JOURNAL CLUB:

Prevalence of Flawed Multiple-Choice Questions in Continuing Medical Education Activities of Major Radiology Journals

David J. DiSantis[1]
Andres R. Ayoob
Lindsay E. Williams

OBJECTIVE. The purpose of this study was to assess whether the continuing medical education (CME) multiple-choice questions (MCQs) in three major radiology journals adhere to standard question-writing principles.

MATERIALS AND METHODS. All CME MCQs (total of 181) in the January 2013 editions of the *AJR*, *RadioGraphics*, and *Radiology* composed the test sample. Each question was evaluated by three reviewers for compliance with seven MCQ-writing guidelines that have been documented in the medical education literature as associated with frequent flaws in medical CME.

RESULTS. Seventy-eight of the 181 (43%) questions contained one to four flaws.

CONCLUSION. A large fraction of radiology CME questions violate standard question-writing principles.

For practitioners and trainees, ongoing maintenance of certification requirements, the advent of exclusively computer-based board examinations, and the ubiquity of continuing medical education (CME)-offering platforms make multiple-choice questions (MCQs) an inescapable part of contemporary radiology and all medical disciplines. As a result demand for well-constructed MCQs has burgeoned. Studies have revealed, however, that examinations from medical school through CME are rife with questions that violate accepted standard question-writing principles [1–6]. In the medical education literature, the term for these nonstandard questions is "flawed" [1–3, 7–9]. Flawed items can be as much as 15 percentage points more difficult to answer correctly than questions that adhere to proper guidelines [3], and tests that include them have a failure rate elevated as much as 25% [2]. In addition, journals' self-assessment CME exercises are now a component of the American Board of Radiology (ABR) Maintenance of Certification program and so should be able to tolerate more rigorous external scrutiny for validity. A seminal 2006 study revealed that all 40 of 40 evaluated CME questions in the *New England Journal of Medicine* contained flaws [1]. We evaluated a substantially larger test group of CME MCQs in three major radiology journals to determine whether they complied with standard question-writing principles.

Materials and Methods

Chance was used to arrive at the CME articles from 1 month in 2013—January—as an unselected sample. The print and online MCQs from that month's editions of the *AJR*, *RadioGraphic*s, and *Radiology* composed the test group. No questions were excluded. The journals offered 181 multiple choice items for their 22 CME-designated articles, as follows: *AJR,* eight articles, 41 questions; *RadioGraphics*, 13 articles, 130 questions; *Radiology,* one article, 10 questions.

Evaluation of the Questions

Haladyna et al. [10] and Rodriguez [11] have validated 31 multiple-choice item-writing guidelines using a meta-analysis of empirical studies and consensus citation in standard educational assessment textbooks. Among those, we targeted seven precepts documented in previous studies as frequently violated in medical education testing, including CME [1–7]. Each of us evaluated each of the 181 questions for compliance with the seven guidelines. Any differences among the assessments were resolved by consensus.

The senior author has completed the question-writing training program conducted by the ABR examination development and psychometrics divisions and served 6 years as chair of an ABR Maintenance of Certification subspecialty examination creation

Keywords: continuing medical education, educational measurement, question writing, radiology education, standards

DOI:10.2214/AJR.13.11963

Received September 11, 2013; accepted after revision August 25, 2014.

[1] All authors: Department of Radiology, University of Kentucky College of Medicine, Chandler Hospital, Rm HX-315A, 800 Rose St, Lexington, KY 40536. Address correspondence to D. J. DiSantis (ddi226@uky.edu).

AJR 2015; 204:698–702

0361–803X/15/2044–698

committee. The second author has completed the question-writing training conducted by the education department at our institution and serves on the examination creation committee of the Association of Medical Student Educators in Radiology.

Descriptions of the Item-Writing Guidelines and Examples of Noncompliant Questions

In discussing MCQs, the following are useful background definitions. The stem is the part of the item that asks for a response. Options are the correct choice, called the key, plus the incorrect choices, called distractors.

Unfocused Stem

According to the ABR *Item Writers' Guide* [12], "The stem should present all the information necessary for the candidate to formulate an answer without having to look for clues in the option list." When the stem is unfocused, the test taker must read all of the options to discern what is being asked. An example of a question with an unfocused stem is, "Which of the following statements about unrepaired transposition of the great vessels is true?" [13]. In our study, 25 questions that dealt with illustrations were not evaluated for stem focus because constructions such as "Which of the following findings is present on the accompanying CT scan?" are appropriate in that setting.

Negative Stem

A negative stem is a flaw whereby the stem includes negatives such as "not" and "except" [1]. According to the ABR guidelines, asking for an answer that is false tests the ability to think conversely rather than knowledge of the subject. Test takers often read this type of stem and mentally proceed to the first correct statement even though it is not the correct test answer [12]. An example of a negative stem question is, "Each of the following statements describes gastric NETs *except*:" [14].

Window Dressing

Questions affected by window dressing contain verbiage irrelevant to the question asked or the concept being assessed [1]. According to the ABR guidelines [12], a stem that "includes too much information not directly related to the question... can be confusing and can draw the candidate away from the task at hand." An example of window dressing is "numbness of the left side of the lower lip of a 45-year-old man with facial trauma due to a motor vehicle collision is most likely caused by a fractured mandible with displacement of the ____" [15]. The superfluous information could be eliminated by condensing the question to "With mandibular fracture, numbness of the lower lip is most likely caused by displacement of the ____."

Unequal Option Length

The ABR guidelines advise that the correct answer be similar in length to the distractors [12]. True statements are usually longer than false statements, making them a potential clue to the correct answer [1]. We counted option length as unequal if one choice was at least twice as long as another, such as "A. Heterogeneous echotexture consistent with hemorrhagic infarct within the testis" and "B. Torsion knot" [16]. We did not apply this guideline to choices of only a few words, such as four versus two words.

Negative Options

Question writers are admonished to phrase choices positively, avoiding negatives such as "not" [10], because "choosing an option in the negative is more a test of reverse thinking than of knowledge" [12]. An example of such an option is "C. Pregnancy is not a risk factor for ovarian torsion" [16].

Clues to the Correct Answer

The unwary question writer can provide tips to the test wise. The following are types of clues.

Vague terms—"Might," "may," and "can" are clues to the correct answer because "they indicate that almost anything is within the realm of possibility" [12], for example, "C. Because of a potential posttreatment tumor dedifferentiation, combined imaging with FDG PET/CT and somatostatin analogs might be useful to follow-up patients" [14].

Specific determiners—Absolute terms, such as "always," "never," and "completely" are clues to incorrect choices because there are no exceptions [10, 12]. An example is "C. The aorta and pulmonary valves are always side by side" [13].

Mutually exclusive pair of options—Savvy test takers may discern that one of a mutually exclusive pair of options [10]—such as "B. Well-defined nonsclerotic margin" and "C. Well-defined sclerotic margin" [17]—is correct (and it usually is) and then have a 50/50 chance of choosing the correct test answer [12]. Two sets of mutually exclusive options would be acceptable, however, because the test-taker gains no statistical advantage [12], for example, "A. Alkylating agent that inhibits osteoclastic activity; B. Alkylating agent that inhibits osteoblastic activity; C. Monoclonal antibody that inhibits osteoclastic activity; D. Monoclonal antibody that inhibits osteoblastic activity" [17].

Grammatical inconsistencies—Grammatical inconsistencies should be avoided [10] so that choices do not stand out because of their phrasing [18], for example [19]:

Delayed liver imaging should be performed:
A. 2–3 minutes after contrast injection.
B. 3–5 minutes after contrast injection.
C. 5–10 minutes after contrast injection.
D. Not less than 10 minutes after contrast injection.

Heterogeneous Options

All options should be in the same general category as the correct answer—for example, all are diagnoses, imaging findings, or treatments [18]. Consequently, question writers should avoid mixed (nonhomologous) options [12] that offer choices that are not similar in content [10], such as:

A. Present in a background of echogenic breast tissue.
B. Over 10 mm in size.
C. Composed of faint, loosely grouped calcifications.
D. Associated with a benign process.

For eight questions (4% of the test pool), we initially varied with one another with regard to whether the options were different enough to be considered heterogeneous. Consensus was reached after comparison with germane examples in the literature [1, 10, 12].

Results

Seventy-eight of the 181 (43%) questions contained flaws. Forty-five questions had one flaw, 24 questions had two, eight questions had three, and one question had four flaws. Specific flaws varied widely in prevalence, as follows: unfocused stem, 39; negative stem, 2; window dressing, 1; unequal option length, 23; negative options, 2; clues to correct answer, 13; heterogeneous options, 38.

Discussion

In comparison with other studies of flawed CME questions in medical journals, our evaluation showed radiology acquitting itself fairly well. A 2006 study, for example, revealed that all 40 of 40 evaluated CME questions in the *New England Journal of Medicine* were flawed [1], and 2007 and 2010 studies of CME questions in the German medical literature [4, 5] revealed flaws in 68% and 65%. Still, 43% of CME questions in leading radiology journals violated standard MCQ item-writing principles.

CME self-assessment testing is a low-stakes proposition, an untimed open-book examination. Test-takers can refer to the article and thereby ferret out the answer to even the most ill-conceived question. And these exercises typically can be retaken as needed. However, the recent designation of self-assessment CME as a key component of the ABR Maintenance of Certification requirements (equivalent to Maintenance of Certification self-assessment modules) has elevated its status and visibility, and hence the likelihood and appropriateness of outside scrutiny. Consequently, concern for the propriety and validity of CME testing is justified. The MCQs should neither violate accepted question-writing tenets of education and testing nor flout the question-writing guidelines and references provided in journals' information for authors material. We should follow our own rules.

Validity, as applied to test scores, refers to the accuracy with which the scores measure a particular ability of interest [21]. Flawed questions can negatively affect test validity [6, 9, 22]. For example, they can be more difficult than questions that adhere to standard guidelines.

Item difficulty refers to the proportion (percentage) of test takers who answer a question correctly [3]. Flawed questions can be up to 15% more difficult than questions in standard form [3]. As may be expected, studies that compare conforming and flawed questions show a 10–25% difference in passing rates, the flawed items failing more test takers [2, 3]. This places some learners at risk of failure for reasons that have nothing to do with their knowledge of the topic. Whereas the focus of our initial study was to assess the prevalence of flawed MCQ items, an intriguing follow-up investigation would be to compare learners' performances on flawed and compliant questions.

The chief limitation of our study was temporal selection and resultant potential for sampling error because we assessed one particular month's CME offerings. Consequently, only one issue of each journal was evaluated. The random pick of a month and the relatively large number (181) of questions evaluated in part mitigate this shortcoming. Because the distribution of question source material was skewed heavily toward *RadioGraphics* (72%, as opposed to 23% for the *AJR* and 5% for *Radiology*), comparison of flaw rates among the journals is not meaningful.

We evaluated the questions using 7 of 31 validated multiple-choice item-writing guidelines [10], specifically focusing on those that have a literature track record of violation in medical education and CME [1–7]. Some guidelines were not applicable (e.g., use of humor if compatible with the learning environment), and others were deemed too subjective (e.g., "Base each item on an important concept to learn"). One reasonably could expect that strict application of all 31 tenets would yield an even higher flawed question rate.

Although it would be ideal if all question-writing guidelines were supported by empirical evidence of their positive effect, research on them has been asystematic. The number of studies reported on has been small relative to the number of guidelines [10]. Although the guidelines may be advice rather than inviolable mandates, the tenets are distilled from the studies available and supplemented by the consensus of educational testing textbook authors with decades of experience [10, 22, 23]. As Downing [23] observed, "these... principles offer the best evidence in practice for creating effective and defensible MCQs."

The medical education literature supports several measures for improving MCQ questions. Results of early work suggest that a web-based application can detect common item flaws, at least partly automating the initial phase of question vetting [6]. As would be expected, formal training sessions for question writers result in higher quality of questions as measured by standard criteria [25, 26]. Given the noncentralized "faculty" writing CME questions, perhaps this could be accomplished via online tutorials supplementing the journals' guidelines for authors.

Incorporation of a committee review process, including experts in not only question writing but also the subject material, likewise yields higher-quality questions [11, 27]. Such vetting bolsters the legitimacy of the evaluation process. According to Haladyna and Rodriguez [28], documentation that items are being reviewed for violations of item-writing rules constitutes one form of validity evidence. The *AJR* has gone a step further in centralizing the CME question-writing process. CME consulting editors, who have subject expertise and experience in item writing, not only choose CME-worthy articles but also create the self-assessment questions themselves [29]. In 2013, *RadioGraphics* added a senior editor for education as part of the quality control process for CME exercises (Ritke R, written communication, 2014). At *Radiology*, editorial and publications staff members review CME exercises for compliance with their standard question guidelines [18, 30] (Humpal J, written communication, 2014).

One valuable source of question validation, that is, assessment of the test takers' performance, is not practical in journal CME. Although performance data could be gathered and potentially used for guiding future examination creation, it is unlikely that CME exercises would be edited once the journal issue had been published.

Brunnquell et al. [6], evaluators of medical education testing, noted, perhaps wryly, that "the typical MCQ item author is a medical expert with time constraints [and] a lack of formal didactic education [in] state-of-the-art item writing principles." That assessment likely describes the hurdles that many radiology CME question authors face. Our results corroborate the hypothesis that as in other medical specialties, a substantial fraction of radiology CME questions violate standard question-writing guidelines. Although self-assessment exercises are not high-stakes examinations, this component of the radiology certification process should be able to bear up under scrutiny.

References

1. Stagnaro-Green AS, Downing SM. Use of flawed multiple-choice items by the *New England Journal of Medicine* for continuing medical education. *Med Teach* 2006; 28:566–568
2. Downing SM. Construct-irrelevant variance and flawed test questions: do multiple-choice item-writing principles make a difference? *Acad Med* 2002; 77:S103–S104
3. Downing SM. The effect of violating standard item writing principles on test and students: the consequences of using flawed test items on achievement examinations in medical education. *Adv Health Sci Educ Theory Pract* 2005; 10:133–143
4. Gutmann A, Degirmenci U, Kreil S, Kornhuber J, Weih M. Improvement of CME questions from *Der Nervenarzt* [in German]. *Nervenzart* 2010; 81:1363–1367
5. Kühne-Eversmann L, Nussbaum C, Reincke M, Fischer MR. CME activities of medical journals: quality of multiple-choice questions as evaluation tool—using the example of the German medical journals *Deutsches Arzteblatt, Deutsche Medizinische Wochenschrift,* and *Der Internist* [in German] *Med Klin (Munich)* 2007; 102:993–1001
6. Brunnquell A, Degirmenci U, Kreil S, Kornhuber J, Weih M. Web-based application to eliminate five contraindicated multiple-choice question practices. *Eval Health Prof* 2011; 34:226–238
7. Tarrant M, Knierim A, Hayes S, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Educ Pract* 2006; 6:354–363
8. National Board of Medical Examiners. Technical

item flaws. In: *Item writing manual,* 3rd ed rev. Philadelphia, PA: National Board of Medical Examiners, 2002. www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf. Accessed September 2, 2013

9. Wadi M. Question vetting: theory and practice. *Educ Med J*. April 2012. www.eduimed.com/index.php/eimj/article/viewFile/29/31. Accessed September 2, 2013
10. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ* 2002; 15:309–334
11. Rodriguez MC. The art and science of item writing: a meta-analysis of multiple-choice format effects. Educational Measurement website. edmeasurement.net/aera/papers/artandscience.pdf. Published April 1997. Revised August 1997. Accessed September 2, 2013
12. American Board of Radiology website. *Item writers' guide*. August 5, 2009. www.aur.org/uploadedFiles/Alliances/AMSER/Educator_Resources/Student_Evaluation/ABR-Item-Writers-Guide.pdf. Accessed September 2, 2013
13. Saremi F, Ho SY, Cabrera JA, Sanchez-Quintana D. Right ventricular outflow tract imaging with CT and MRI. Part 1. Morphology. *AJR* 2013; 200:[web]W39–W50
14. Sahani DV, Bonaffini PA, Fernandez-DelCastillo C, Blake MA. Gastroenteropancreatic neuroendocrine tumors: role of imaging in diagnosis and management. *Radiology* 2013; 266:38–61
15. Winegar BA, Murillo H, Tantiwongkosi B. Spectrum of critical imaging findings in complex facial skeletal trauma. *RadioGraphics* 2013; 33:3–19
16. Lubner MG, Simard ML, Peterson CM, Bhalla S, Pickhardt PJ, Menias CO. Emergent and nonemergent nonbowel torsion: spectrum of imaging and clinical findings. *RadioGraphics* 2013; 33:155–173
17. Chakarun CJ, Forrester DM, Gottsegen CJ, Patel DB, White EA, Matcuk GR Jr. Giant cell tumor of bone: review, mimics, and new developments in treatment. *RadioGraphics* 2013; 33:197–211
18. Collins J. Education techniques for lifelong learning: writing multiple choice questions for continuing medical education activities and self-assessment modules. *RadioGraphics* 2006; 26:543–551
19. Liu YI, Shin LK, Jeffrey RB, Kamaya A. Quantitatively defining washout in hepatocellular carcinoma: self-assessment module. *AJR* 2013; 200:84–89
20. Wang LC, Sullivan M, Du H, Feldman MI, Mendelson EB. US appearance of ductal carcinoma in situ. *RadioGraphics* 2013; 33:213–228
21. Ebel RL, Frisbie DA. *Essentials of educational measurement*, 5th ed. Englewood Cliffs, NJ: Prentice Hall, 1991
22. Downing SM, Haladyna TM. Validity and its threats. In: Downing SM, Yudkowsky R, eds. *Assessment in health professions education*. New York: Routledge, 2009:21–56
23. Haladyna TM. *Developing and validating multiple choice test items*, 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 1999
24. Downing SM. Written tests: constructed-response and selected-response formats. In: Downing SM, Yudkowsky R, eds. *Assessment in health professions education*. New York: Routledge, 2009:149–184
25. Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. *Adv Health Sci Educ Theory Pract* 2012; 17:369–376
26. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew R. The quality of in-house medical school examinations. *Acad Med* 2002; 77:156–161
27. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a committee review process to improve the quality of course examinations. *Adv Health Sci Educ Theory Pract* 2006; 11:61–68
28. Haladyna TM, Rodriguez MC. *Developing and validating test items*. New York: Routledge, 2013
29. CME consulting editors bring additional expertise to *AJR* CME activities. *AJR Editorial Update* 2013; 3:8. www.arrs.org/uploadedFiles/EIC_Newsletter.pdf. Accessed September 2, 2013
30. DiSantis DJ. Writing good multiple-choice questions: a brief guide for radiologists. *RadioGraphics* 2013; 33:1865–1866

FOR YOUR INFORMATION

This article has been selected for *AJR* Journal Club activity. The accompanying Journal Club study guide can be found on the following page.

APPENDIX 1: *AJR* JOURNAL CLUB

Study Guide

Prevalence of Flawed Multiple-Choice Questions in Continuing Medical Education Activities of Major Radiology Journals

Margaret Mulligan[1], Joseph J. Budovec[1], Alan Mautz[2]

[1]Medical College of Wisconsin, Milwaukee, WI

[2]The Aroostook Medical Center, Presque Isle, ME

*mmulliga@mcw.edu, jbudovec@mcw.edu, alan.mautz@emhs.org**

Introduction

1. Is this study relevant and timely?
2. How would you formally state the research question? How would you state the hypothesis and alternative hypotheses?

Methods

3. How were continuing medical education questions selected for assessment in the study?
4. What were the inclusion criteria for questions and journals in the study? What were the exclusion criteria for questions and journals in the study?
5. Does a 1-month unselected sample establish an adequate sample size for the study? What biases may arise from the selection criteria used to create the study sample?
6. How did the study establish the standards, or guidelines, for evaluating the continuing medical education questions?
7. What are the qualifications of the three individuals completing the analysis of the questions?
8. Was the method of resolving discrepancies adequately described?
9. In this type of research study, what are standard methodologies for resolving discrepancies?
10. What bias could be introduced by the method of discrepancy resolution used in this study?
11. What were the limitations of this study? Were these limitations adequately addressed?
12. What precepts or variables were selected for analysis? Were the parameters of the variables fully described? In designing a similar study, would you have selected the same variables? Would you be able to repeat the study?
13. What statistical methods were used in the study analysis? The study notes there were 31 validated multiple choice item-writing guidelines, but only seven were selected for focus in the study. What was the justification for the selection of seven versus all 31?

Results

14. Was the study question(s) answered? Were the hypotheses resolved?
15. Was the study designed appropriately, and was the sample size large enough to draw conclusions about the impact of flawed continuing medical education questions in radiology journals?

Discussion

16. What standards exist for writing questions for continuing medical education activities in the medical literature? Is there a different standard for radiology?
17. What are the implications of flawed continuing medical education questions? Do users fail to reach a correct answer?
18. Is the study adequately designed and powerful enough to influence radiologic publications to raise standards related to writing continuing medical education questions?

Background Reading

1. Stagnaro-Green AS, Downing SM. Use of flawed multiple-choice items by the *New England Journal of Medicine* for continuing medical education. *Med Teach* 2006; 28:566–568
2. Collins J. Education techniques for lifelong learning: writing multiple choice questions for continuing medical education activities and self-assessment modules. *RadioGraphics* 2006; 26:543–551

*Please note that the authors of the Study Guide are distinct from those of the companion article.