Title
Does Educator Training or Experience Affect the Quality of Multiple-Choice Questions?

Journal
Academic Radiology, 22(10)

ISSN
1076-6332

Authors
Webb, Emily M
Phuong, Jonathan S
Naeger, David M

Publication Date
2015-10-01

DOI
10.1016/j.acra.2015.06.012

Peer reviewed

Does Educator Training or Experience Affect the Quality of Multiple-Choice Questions?

Emily M. Webb, MD, Jonathan S. Phuong, BS, David M. Naeger, MD

Rationale and Objectives: Physicians receive little training on proper multiple-choice question (MCQ) writing methods. Well-constructed MCQs follow rules, which ensure that a question tests what it is intended to test. Questions that break these are described as "flawed." We examined whether the prevalence of flawed questions differed significantly between those with or without prior training in question writing and between those with different levels of educator experience.

Materials and Methods: We assessed 200 unedited MCQs from a question bank for our senior medical student radiology elective: an equal number of questions (50) were written by faculty with previous training in MCQ writing, other faculty, residents, and medical students. Questions were scored independently by two readers for the presence of 11 distinct flaws described in the literature.

Results: Questions written by faculty with MCQ writing training had significantly fewer errors: mean 0.4 errors per question compared to a mean of 1.5–1.7 errors per question for the other groups ($P < .001$). There were no significant differences in the total number of errors between the untrained faculty, residents, and students (P values .35–.91). Among trained faculty 17/50 questions (34%) were flawed, whereas other faculty wrote 38/50 (76%) flawed questions, residents 37/50 (74%), and students 44/50 (88%). Trained question writers' higher performance was mainly manifest in the reduced frequency of five specific errors.

Conclusions: Faculty with training in effective MCQ writing made fewer errors in MCQ construction. Educator experience alone had no effect on the frequency of flaws; faculty without dedicated training, residents, and students performed similarly.

Key Words: Multiple-choice questions; educator experience; question flaws; education.

Physicians are rarely trained to properly write multiple-choice examinations, including those working in academic settings. However, this skill set has become much more relevant in recent years. With the transition to the new written format of radiology board certification examinations (1), the development of more rigorous self-assessment requirements for maintenance of certification examinations (2–4), and the greater inclusion of radiology into integrated medical student curricula (5), multiple choice radiology questions are in great demand.

Well-constructed multiple-choice questions (MCQs) follow a set of parameters that ensure the question tests what it is intended to test (6–8). Questions that violate widely agreed on rules are described in the education literature as flawed (9–13). In simple terms, a flawed question tends to test "how good of a test taker" someone is, rather than the relevant knowledge intended, which can disadvantage some students (10). Previous literature examining MCQs has revealed that such mistakes are common within continuing medical education (CME) materials (14,15) and on health care sciences examinations (10,16).

Previous authors have found that MCQ writing is improved after dedicated faculty training (17). However, to our knowledge, there has been no previous assessment as to whether educator experience level otherwise affects the quality of MCQs.

We sought to determine how often MCQ writing rules were violated among questions submitted for use in our primary medical student radiology elective, and whether the prevalence of flawed questions differed significantly among question writers with varying levels of experience and training. Our hypothesis was that faculty with prior training in question writing would perform best, and that of those without training, educators with those most experience (faculty > residents > students) would perform better than those with less experience.

MATERIALS AND METHODS

Study Design

At our institution, we have access to a large bank of internally generated MCQs that were previously submitted for use in

Acad Radiol 2015; 22:1317–1322

From the Department of Radiology and Biomedical Imaging, University of California, San Francisco, 505 Parnassus Avenue, M-391, San Francisco, CA 94143-0628 (E.M.W., J.S.P., D.M.N). Received May 28, 2015; accepted June 27, 2015. **Address correspondence to:** E.M.W. e-mail: emily.webb@ucsf.edu

http://dx.doi.org/10.1016/j.acra.2015.06.012

our main medical student radiology elective by (a) core education faculty in the Department of Radiology who have had dedicated training in question writing, (b) other radiology faculty, (c) third and fourth year radiology residents, and (d) fourth year medical students. "Core education faculty" was defined as those instructors who held leadership positions in departmental educational endeavors and had previous instruction in MCQ writing. Three faculty members, with subspecialty training in nuclear medicine, cardiopulmonary, and abdominal imaging, met this description. One core faculty member was the medical student elective course director and Co-Director of Medical Student Education for the Department of Radiology, second core faculty member was the curriculum steward for radiology for the School of Medicine and Co-Director of Medical Student Education for the Department of Radiology, and the third was the Director of CME for the Department of Radiology. The three core instructors had 4, 9, and 9 years of faculty educator experience, respectively, whereas general faculty was comparatively comprising educators with a wide range of experience from 1 to >30 years. Previous training in question writing among the core education faculty varied. One core faculty member completed an institutional 2-hour faculty development workshop on MCQ writing that was intended for medical school teaching faculty. One completed a 1-hour WebEx training session on MCQ writing provided by the American Board of Radiology. The third received 2 hours of one-on-one mentoring in MCQ writing from medical school faculty as part of a junior faculty mentoring program. All three core faculty also previously completed independent review of published MCQ writing resources (18–20).

Questions used on our graded radiology course examinations are all vetted and edited by the course directors. The question bank, however, also includes the initial, unedited versions of all questions submitted for consideration for the course (n = approximately 400). Unedited questions are categorized into separate files by the rank of the author (faculty, resident, student, and so forth), but have no other identifying author information. Given that the questions written by core education faculty had to be parsed out from other faculty submissions, the radiology medical student education coordinator, who was not otherwise involved in the study, separated the faculty questions into two different files, based on email records of initial question submission. Given overlapping topics and rotation in course instructors, the other question writers could not reasonably be identified by the questions alone. Questions covered all radiology subspecialty areas. Topic areas were not specific to educator level.

Relatively fewer resident written items were available in the question bank (approximately 50), so 50 unedited questions from each of these four educator groups were randomly selected. The questions were assembled together in a single 200 question PowerPoint presentation, with the question sequence randomized. A key to the "educator level" of each question writer was maintained by one author who did not evaluate the questions. The PowerPoint formatting of each question slide was identical.

The core education faculty who contributed questions consisted of three writers. The questions from the other groups (noncore faculty, residents, and students) were selected consecutively and anonymously, so the exact number of distinct authors in each group is unknown. However, the number of authors contributing to the question bank was >20 for each of these three categories.

MCQ Flaws

The list of question flaws was generated by a literature review (11–13,16), referencing a local institutional guideline for MCQ writing (18), review of the American Board of Radiology item writers' guide (19), and the National Board of Medical Examiners' item writing guide (20). All commonly cited independent "flaws," which violate standard evidence-based principles of effective item writing, were included in the analysis. "Repeat wording" in answer options was omitted as an independent flaw as it overlaps considerably with the concept of "element repetition," which was included in a separate rule. It was felt that all questions exhibiting repeat wording would be identified through the more general rule. Additionally, some recommendations including organization of the sequence of answer options (alphabetized vs. logical order) were not included as there are conflicting recommendations (18,19).

Eleven common MCQ writing flaws were included and defined as follows:

1. Content not important
   Examination questions should be based on concepts that are important for the learners to "take away" from the session. In other words, they should focus on main points, not minutia.
2. Open-ended or unfocused stems
   MCQs have two components: the "stem" and the answer choices. The stem is the question or incomplete statement. Most of the information pertinent to the question should be in the stem, so that a test taker can reasonably determine the correct answer before even reading the possible answer choices. However, when a stem is unfocused, the test taker must read all the options before they can determine what the question is asking. An example of an open-ended stem is, "Nuclear medicine tests are…."
3. Negative stem or negative answer options
   Question stems should avoid negative phrasing such as, "Which of these is NOT," "…EXCEPT," "…FALSE," and so forth. Negative wording in answer options should be avoided as well. Negative terms often make a question unnecessarily confusing.
4. Inclusion of superfluous information
   The goal is to present the information in a manner that is concise and uncomplicated. Avoiding superfluous

information that is irrelevant to the question being asked can be unnecessarily confusing.

5. Use of multiple answer options
Questions should be single best answer. Those with an "all of the above" option can make the question easier. If the test taker recognizes that any answer choice is incorrect they can also eliminate "all of the above" as the correct choice. Likewise, if the test taker recognizes that any two choices are correct then "all of the above" must be the correct answer. Use of "none of the above" makes a question more difficult.

6. Too many or too few answer options
In general, there should be four to five answer options per question. Although in some cases fewer answer options may be acceptable if there are not enough reasonable, logical distractors to include four to five options.

7. Unequal option length
Experienced test takers will often pick the longest answer option as correct. Correct answers often require more information to make them correct. To avoid this clue, all answers should be similar in length. As this is somewhat subjective, for this study, we defined "similar in length" to mean that no answer option was more than twice as long as another. We excluded answer options that were only one or two words, as their inclusion made the rule too sensitive. For example, if answer options of only a few words were considered, an answer might be deemed "too long" simply because one disease entity has a longer name than the other options. This definition has been used by previous authors (14).

8. Use of absolute or vague terms
Answer choices with absolute terms such as "always," "never," and "all" are often false and an experienced test taker will exclude them automatically. Vague terms such as "probably" and "usually" should also be avoided as they are open to subjective interpretation.

9. Inclusion of grammatical clues
Grammatical clues occur when all answer options do not follow grammatically from the stem. For example, the most sensitive test for gastric cancer is *an*… (A) CT scan, (B) MR, (C) Endoscopy, (D) Double contrast upper GI. The word "an" in the stem implies that endoscopy is the correct answer as it is the only answer option to begin with a vowel.

10. Inclusion of logical clues (mutually exclusive or exhaustive pair options)
A subset of answers can sometimes be mutually exclusive or exhaustive. For example, "When the ACR appropriateness criteria are used, what is the effect on physician ordering practices? (A) More imaging studies are ordered, (B) the same number of imaging tests are ordered, (C) fewer imaging studies are ordered, (D) more MR studies are ordered, (E) fewer contrast enhanced studies are ordered." In the preceding question, options A, B, and C are exhaustive. One of those options must be the correct answer.

11. Similarity of option elements (convergence)
Sometimes referred to as "convergence strategy," the correct answer for any question is more likely to have elements in common with the other answer options. Thus, answer options with repeating elements give a clue to the correct answer. And likewise, answer options or distractors that are outliers are unlikely to be correct. For example, "Which imagining study is best to diagnose nephrolithiasis? (A) A supine noncontrast CT, (B) a prone noncontrast CT, (C) a CT with contrast, (D) a plain film." In the preceding example, options A and B have the most elements in common and one of those is most likely to be correct. Option D is an outlier, the only option that does not suggest CT, and therefore is the least likely to be correct.

Data Collection

Two readers who were blinded to the identities and experience level of the question writers independently scored each question for the presence or absence of these 11 distinct flaws, which violate standard evidence-based principles of effective item writing. Items that were scored differently between the two readers were examined again by both and scored by consensus.

Statistical Analysis

Overall differences in error frequency among the four groups were assessed with the Kruskal–Wallis equality-of-populations rank test. Tests evaluating for pairwise differences were achieved with the Wilcoxon rank sum test. Comparisons for differences in breaking each individual rule among the four groups were achieved using the Kruskal–Wallis equality-of-populations rank test, although pairwise comparison was not made for each rule to avoid excessive multiple comparisons (66 pairwise tests to evaluate 11 rules), and because of the low statistical power (multiple rules had relatively few errors). P values $<.05$ were considered statistically significant. With 50 questions per group, we were powered to be able to detect pairwise differences of approximately 0.4 errors per question between any two groups (beta = 0.8).

RESULTS

Two hundred examination questions were analyzed for the presence of 11 flaws resulting in 2200 yes or no evaluations; there was initial disagreement between reviewers on only 17 of 2200 items (0.8%). These were reviewed again by both readers in a consensus fashion with no residual disagreements.

Of 200 questions, 136 (68%) contained at least one flaw (range of flaws, 1–4). Among core education faculty, only 17/50 (34%) contained at least one flaw and no question

had more than two flaws. Among the three other groups flaws were more common. Noncore faculty had 38/50 (76%) questions with at least one flaw (range 1–4), residents had 37/50 (74%) questions with at least one flaw (range 1–4), and students had 44/50 (88%) questions with at least one flaw (range 1–4).

Overall, 254 flaws were identified across 200 questions for a mean of 1.3 errors per question. Questions written by core education faculty had significantly fewer errors: mean = 0.4 errors per question compared to a mean of 1.5–1.7 errors per question for the other groups ($P < .001$) (Table 1). Pairwise statistical comparisons showed a statistical significant difference between core education faculty and all other groups (all P values <.001), with no statistically significant differences among the other three groups (P values range from .35–.91).

The distribution of errors across the different groups is summarized in Table 2. Of note, for 6/11 flaws, the frequency of the flaw differed among the groups, in most cases because the core education faculty performed better than the other groups (flaws 1–6, Table 2). In one case, the students performed worse than the other groups (flaw 6, Table 2). For 5/11 flaws there were no statistically significant differences, in some cases because few errors were made in any group (flaws 8, 9, and 10, Table 2) and in two cases because the errors occurred in all groups at approximately the same frequency (flaws 7 and 11, Table 2).

TABLE 1. Prevalence of Multiple Choice Question Flaws for Different Educator Groups

Group	Number of Questions	Number of Flaws Overall	Number of Flawed Questions	Mean Errors per Question
Trained Faculty	50	19	17 (34%)	0.4
Faculty	50	74	38 (76%)	1.5
Residents	50	76	37 (74%)	1.5
Students	50	85	44 (88%)	1.7
Totals	200	254	136 (68%)	1.3

DISCUSSION

There is likely some assumption in academic medicine that academic physicians would inherently understand how to write well-constructed test questions. This assumption probably stems from recognition of the large number of tests physicians take during their training. Also, implied competence may derive from academic physicians successfully performing other "academic" tasks, such as writing articles and lecturing. Many of these academic tasks have been traditionally taught by the "see one, do one" model rather than formal pedagogy. For these reasons, dedicated training in question writing is often not offered or is only cursorily taught.

Our study found that faculty who had some previous instruction in proper question-writing techniques made fewer errors than other groups of educators. This may seem intuitive. However, untrained faculty, who committed flaws far more than trained faculty, actually faired no better than medical students. In fact, untrained faculty, residents, and students all performed similarly. Previous authors have also demonstrated that MCQ writing improves after targeted training workshops (20).

There were five flaws in question writing in particular that trained faculty made less frequently than the other groups:

- Pertinent, important concepts. This may be only partially attributable to training in question writing, and may also be in part because of the core education faculty being more aware of the overall course goals and student learning objectives.
- Unfocused or open stems.
- Negative phrasing.
- Multiple answer options.
- Suboptimal number of answer options.

In fact, among these five rules, trained faculty only made one error on one question (negative phrasing, $n = 1$). It is also notable that it was among these flaws where the largest number of errors was made overall (a total of 165 errors for these five flaws, of which only 1, or 0.6% was made by a trained faculty). The single most common flaw, an unfocused stem, which accounted for 57/254 or 22% of all errors, was not made once by trained faculty. Therefore, this group of flaws represents an area where training can make significant improvements in test item quality.

Conversely, there were two item flaws that even faculty trained in question writing still made with relative frequency, similar to question writers with less experience, including:

- answer options of differing length;
- similarity of option elements (convergence).

All groups made the differing length error with moderate frequency. This may be in part because of the subjective nature of this question-writing rule. We used a definition selected by previous authors (14): no answer item should be more than twice the length of another (excluding those that were only one or two words). However, there is no similar published definition in question writing manuals. It is likely that the question writers who were aware of the rule, at least in some cases, might have felt their answer choices were *similar enough* in length.

The second was the use of answer items that were similar in some way, where the correct answer includes the most elements in common with the other options. Convergence, as it is sometimes called, is known to be one of the more subtle but common errors in MCQs (20). This flaw may occur if a question writer begins with the correct answer option and then tries to generate variations of the correct answer as distractors. Therefore, answer options that are outliers (and do not share features in common with other options) are less

TABLE 2. Distribution of Each Multiple Choice Question Flaw Across Different Educator Groups

	Number of Occurrences					
Question Flaw	Core Faculty	Faculty	Residents	Students	Totals	Statistical Difference, *P* Values*
(1) Content not important	0	4	7	11	22	*.004*
(2) Unfocused stem	0	20	22	15	57	*.001*
(3) Negative phrasing	1	15	16	8	40	*.001*
(4) Multiple answer options	0	9	7	9	25	*.019*
(5) Too many or few answer options	0	9	4	8	21	*.013*
(6) Superfluous information	0	1	0	11	12	*.001*
(7) Unequal option length	6	4	7	11	28	.231
(8) Absolute or vague terms	0	3	5	3	11	.181
(9) Grammatical clues	2	2	0	0	4	.255
(10) Logical clues	0	1	1	3	5	.275
(11) Convergence	10	6	7	6	29	.631
Totals	19	74	76	85	254	*.0001*

*Note: Italicized *P* values are statistically significant.

likely to be correct. Of course, this mental sequence is common in question writing. We first decide what we want to test, already know the correct answer, and then come up with 3–4 reasonable, plausible distractors. If distractors are not plausible, they would not be distracting. Writing good distractors can be one of the most challenging elements of question writing, and it is therefore somewhat unsurprising that convergence remains a common flaw even among seasoned educators. Awareness that this remains a common pitfall may help improve adherence to this MCQ writing standard.

Among the three groups without training in question writing, there was only one flaw where we found a significant difference in prevalence. Most notably, student written questions were far more likely to include potentially confusing superfluous information. Some of these instances were because of inclusion of long clinical vignettes in the question stem. This practice is known to be common in medical school examinations at our institution, and they were likely mimicking this format because of familiarity. Inclusion of extraneous information in the setting of a clinical vignette is more allowable, with some authors (13) suggesting it should not be considered a flaw. Questions with longer clinical vignettes, which often describe both relevant and irrelevant details of a patient presentation, have been shown to be more difficult than shorter vignettes or nonvignette style questions; however, they are not more effective at discriminating those who know the content from those who do not (13). The style still has its proponents although, who believe that it more closely mimics the reality of clinical practice, is a more authentic test of clinical knowledge (13), and can make the examination more interesting for the student (18). In fact, this question format is both used and endorsed by the National Board of Medical Examiners (20).

Lastly, there were several flaws that were uncommon and were not made frequently by any group. These included use of absolute or vague terms, use of grammatical clues in the stem, and use of logical clues (for example, a subset of answers being mutually exhaustive or mutually exclusive). It is possible that these flaws are more recognizable even to the untrained question writer, and are probably more a result of carelessness than lack of training.

The rates of flawed questions in our study are similar to those previously demonstrated in the literature, where the range has been reported as 43%–100% in studies assessing MCQs for CME (14,15). Of note, no previous studies assessing these question-writing rules have found flaw rates as low as 34%, the rate for faculty with previous training in our study, which speaks to both the importance of formal training in question writing and the potential room for improvement that may come from such targeted education. One prior study reports a much lower prevalence of item flaws (approximately 20% before MCQ writing training and only 2% after); however, the methodology does not describe which or how many flaws were assessed (17).

Our study has several limitations. First, our identified list of MCQ flaws is not exhaustive. It simply represents a summary of the most commonly cited flaws. Nevertheless, it is very similar to the set of flaws examined by other authors in studies evaluating adherence of MCQs to item writing standards (14,15). Second, we did not include an assessment as to what level of knowledge (Bloom's taxonomy) (21) was tested by each question. Valid examinations should include questions that measure different levels of knowledge, with careful attention to include those at higher levels, requiring application of knowledge and analysis, for example. However, the lack of such a distribution is a flaw with the examination overall, not with any particular question, which may be valid regardless of Bloom's taxonomy level. Furthermore, there was heterogeneity among the group of core faculty educators as to their previous training in question writing. It is possible that some were

well aware of all 11 of these question flaws, and some may have had less robust training. However, given that errors among the core faculty were so comparatively rare, this may suggest that any amount of training may be beneficial. Because faculty in the noncore group was deidentified, individuals were not surveyed as to any past independent training in question writing. It is certainly possible that some may have independently reviewed MCQ writing materials in the past. Nevertheless, their performance was no better than residents or medical students. Additionally, the group of core faculty educators comprised a much smaller group of question writers, thereby this group may not be broadly representative off all core faculty educators at other institutions. Lastly, all subjects were from a single institution, which limits the generalizability of these findings to a larger population.

CONCLUSIONS

We found that faculty who had been trained in the evidence-based principles of effective MCQ writing made fewer technical mistakes in item construction. However, beyond that, educator experience had no effect on the frequency of flaws with faculty, residents, and students performing similarly. Furthermore, although some question flaws were pervasive, the most commonly encountered mistakes were also shown to be the most correctable with training. Therefore, education efforts in this area will likely have a substantial positive effect on the effectiveness and validity of radiology examinations.

REFERENCES

1. Initial Certification: Diagnostic Radiology. Available at: http://www.theabr.org//ic-dr-landing. Accessed May 18, 2015.
2. Madewell JE, Hattery RR, Thomas SR, et al. American Board of Radiology: maintenance of certification. Radiology 2005; 234:17–25.
3. ABMS Evidence Library. Available at: http://www.abms.org/initiatives/committing-to-physician-quality-improvement/evidence-library/. Accessed May 20, 2015.
4. Maintenance of Certification. Available at: http://www.theabr.org/moc-gen-landing. Accessed May 20, 2015.
5. Straus CM, Webb EM, Kondo KL, et al. Medical student radiology education: summary and recommendations from a national survey of medical school and radiology department leadership. J Am Coll Radiol 2014; 11: 606–610. http://dx.doi.org/10.1016/j.jacr.2014.01.012. Epub Apr 6, 2014.
6. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. Med Teach 2004; 26:709–712.
7. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ 2002; 15:309–334.
8. 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.
9. Downing SM. Construct-irrelevant variance and flawed test questions: do multiple-choice item-writing principles make a difference? Acad Med 2002; 77:S103–S104.
10. 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.
11. Brunnquell A, Degirmenci U, Kreil S, et al. Web-based application to eliminate five contraindicated multiple-choice question practices. Eval Health Prof 2011; 34:226–238.
12. Case SM. The use of imprecise terms in examination questions: how frequent is frequently? Acad Med 1994; 69(10 Suppl):S4–S6.
13. Case SM, Swanson DB, Becker DF. Verbosity, window dressing, and red herrings: do they make a better test item? Acad Med 1996; 71(10 Suppl): S28–S30.
14. DiSantis DJ, Ayoob AR, Williams LE. Journal club: prevalence of flawed multiple-choice questions in continuing medical education activities of major radiology journals. AJR Am J Roentgenol 2015; 204:698–702. http://dx.doi.org/10.2214/AJR.13.11963.
15. 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.
16. Tarrant M, Knierim A, Hayes SK, et al. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Educ Pract 2006; 6:354–363. http://dx.doi.org/10.1016/j.nepr.2006.07.002. Epub Nov 13, 2006.
17. Abdulghani HM, Ahmad F, Irshad M, et al. Faculty development programs improve the quality of multiple choice questions items' writing. Sci Rep 2015; 5:9556. http://dx.doi.org/10.1038/srep09556.
18. Masters S. Tips for constructing multiple choice exam questions [faculty development handout]. San Francisco, CA: School of Medicine, University of California, San Francisco, 2010.
19. American Board of Radiology website. Item writers' guide. www.aur.org/uploadedFiles/Alliances/AMSER/Educator_Resources/Student_Evaluation/ABR-Item-Writers-Guide.pdf; 2009. Accessed September 2, 2013.
20. National Board of Medical Examiners. Technical item flaws. In: Item writing manual. 3rd ed. Philadelphia, PA: National Board of Medical Examiners. www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf; 2002. Accessed September 2, 2013.
21. Bloom BS. Taxonomy of educational objectives: cognitive domain. New York: Longman, 1984.