RadioGraphics

Education Techniques for Lifelong Learning

Writing Multiple-Choice Questions for Continuing Medical Education Activities and Self-Assessment Modules[1]

Jannette Collins, MD, MEd

TEACHING POINTS
See last page

The multiple-choice question (MCQ) is the most commonly used type of test item in radiologic graduate medical and continuing medical education examinations. Now that radiologists are participating in the maintenance of certification process, there is an increased need for self-assessment modules that include MCQs and persons with test item-writing skills to develop such modules. Although principles of effective test item writing have been documented, violations of these principles are common in medical education. Guidelines for test construction are related to development of educational objectives, defining levels of learning for each objective, and writing effective MCQs that test that learning. Educational objectives should be written in observable, behavioral terms that allow for an accurate assessment of whether the learner has achieved the objectives. Learning occurs at many levels, from simple recall to problem solving. The educational objectives and the MCQs that accompany them should target all levels of learning appropriate for the given content. Characteristics of effective MCQs can be described in terms of the overall item, the stem, and the options. Flawed MCQs interfere with accurate and meaningful interpretation of test scores and negatively affect student pass rates. Therefore, to develop reliable and valid tests, items must be constructed that are free of such flaws. The article provides an overview of established guidelines for writing effective MCQs, a discussion of writing appropriate educational objectives and MCQs that match those objectives, and a brief review of item analysis.

©RSNA, 2006

Introduction

The multiple-choice question (MCQ) is the most common type of written test item used in undergraduate, graduate, and postgraduate medical education (1). MCQs can be used to assess a broad range of learner knowledge in a short period of time. Because a large number of MCQs can be developed for a given content area, which provides a broad coverage of concepts that can be tested consistently, the MCQ format allows for test reliability. If MCQs are drawn from a representative sample of content areas that constitute predetermined learning outcomes, they allow for a high degree of test validity. Critics of MCQs argue that higher-level learning cannot be tested with MCQs. However, this criticism is more often attributed to flaws in the construction of the test items rather than to their inherent weakness. Appropriately constructed MCQs result in objective testing that can measure knowledge, comprehension, application, and analysis (2). Disadvantages of MCQs are that they test recognition (choosing an answer) rather than recall (constructing an answer), they allow for guessing, and they are difficult and time-consuming to construct.

The principles of writing effective MCQs are well documented in educational measurement textbooks, the research literature, and test-item construction manuals designed for medical educators (3–5). Yet, a recent study from the National Board of Medical Examiners showed that violations of the most basic item-writing principles are very common in medical education tests (6).

The number of radiologists who will be writing MCQs is expected to increase as more radiologists develop self-assessment modules (SAMs) for the American Board of Radiology's maintenance of certification (MOC) program. In a 10-year period, enrollees in MOC must complete 20 SAMs that include MCQs (7). All diplomates certified in 2002 and beyond are automatically enrolled in the MOC program, and the ABR is encouraging all diplomates to enroll in MOC.

MCQs are difficult and time-consuming to construct, even for individuals who have been formally trained in their construction. Professional test-item writers plan on 1 hour or more to write one good item (8). This article provides guidelines that can be used by radiologists in writing MCQs for SAMs and other continuing medical education materials, as well as for medical student clerkship tests, radiology resident in-service examinations, and written board examinations. Three areas are addressed: *(a)* writing educational objectives, *(b)* defining levels of learning for each objective, and *(c)* writing effective MCQs to test that learning. In addition, test-item analysis is briefly discussed.

Immeasurable learning objective	Understand the appearance of pneumothorax on a supine chest radiograph
Measurable learning objective	Describe five findings of pneumothorax that can be seen on a supine chest radiograph

Figure 1.   Examples of immeasurable and measurable objectives are given. In the immeasurable objective, it is not clear how the student will show that he or she "understands." In comparison, with the measurable objective, it is clear how the student will demonstrate learning, and the qualifier of "five" indicates a specific level of knowledge.

Writing Educational Objectives and Defining Levels of Learning

Good test question writing begins with identifying the most important information or skill that is to be learned. A direct relationship between instructional objectives and test items must exist. Thus, test items should come directly from the objectives (2) and focus on important, relevant content; this emphasis helps test writers avoid testing the knowledge of medical trivia. Controversial test items should be avoided, especially when the knowledge is incomplete or the facts are debated (9). Determining the appropriate test questions can be facilitated by reviewing the major subtopics of the article or other content and by identifying sentences that summarize main ideas or principles. From this information, key facts can be written as declarative sentences, creating a clear picture of what the student should learn. It has been suggested that if the idea, when it is written as an explicit statement, proposition, or principle, forms an important part of the instruction, it is worth testing (10).

Objectives should be written in terms of specific learner behavior and not what the program will teach. They should define important knowledge or skills and should be supported by the instruction provided through the educational program. Observable, measurable objectives allow for accurate assessment of whether the learner has

RadioGraphics

Examples of Objectives and MCQs for Three Levels of Learning		
Level	Objective	Question
Knowledge (learner must recall memorized information but not explain or apply it)	State the average effective radiation dose from chest CT	What is the average effective radiation dose from chest CT? a. 1 mSv b. 8 mSv c. 16 mSv d. 24 mSv
Combined comprehension and application (learner must demonstrate an ability to use, not just explain, new information, applying rules, methods, concepts, principles, laws or theories)	Compare the radiation exposures from different radiologic examinations	Which of the following imaging examinations is associated with the highest effective radiation dose? a. Abdominal and pelvic multidetector CT b. Coronary artery multidetector CT c. Conventional pulmonary angiography d. Digital pulmonary angiography
Problem solving (learner must understand components of a concept, and their relationships to each other, and analyze information)	Explain the effects that various factors have on radiation dose from chest CT	Which of the following actions would decrease the radiation dose from chest CT the least? a. Decreasing mA from 250 to 125 b. Decreasing kVp from 140 to 120 c. Decreasing the pitch from 2 to 1 d. Decreasing scan time from 1 to 0.5

achieved an objective. Examples of measurable terms are *state, explain, list, identify,* and *compare.* Immeasurable terms include *know, understand, learn,* or *become familiar with.* Figure 1 illustrates the difference between an immeasurable objective and a measurable one.

In 1959, Bloom (11) published a taxonomy of cognitive learning, which was described as a hierarchy of knowledge, comprehension, application, analysis, synthesis, and evaluation. Educators have adopted Bloom's taxonomy for test development (12,13), and some have simplified and collapsed it into three general levels (14). The three levels include the following categories: *(a)* knowledge (recall or recognition of specific information), *(b)* combined comprehension and application (understanding or being able to explain in one's own words previously learned information and using new information, rules, methods, concepts, principles, laws, and theories), and *(c)* problem solving (transferring existing knowledge and skills to new situations). A MCQ should test at the same level of learning as the objective it is designed to assess. The Table shows examples of MCQs and objectives for each level of learning.

If the desired outcome of an educational program involves having participants do more than recall facts, the program should be designed to enable learners to apply knowledge or skills. The program's objectives and test questions should reflect different levels of learning. Thoughtfully written objectives are critical to the construction of appropriate test questions and in ensuring adequate assessment of intended learner competence. MCQs written to test knowledge (lower-level learning) would not be appropriate to test competence for objectives that reflect comprehension (higher-level learning). For example, a MCQ that asks the learner to recognize benign dermal calcifications on a mammogram does not test the learner's problem-solving ability. A question that provides specific patient information and imaging data (ie, a patient vignette) and that asks the learner to choose the most appropriate management is an example of an item that tests problem-solving ability.

Test items composed of patient vignettes offer several benefits in addition to assessing application of knowledge. Because such questions require problem solving, they increase the validity of the examination. Such items are more likely to focus on important information, rather than on trivia. Lastly, they help identify examinees who have memorized facts but are unable to use the information effectively.

Teaching Point

Phrase repeated in each option	Which of the following actions would decrease radiation dose by one-half? a. Decreasing mA by one-quarter. b. Decreasing mA by one-third. c. Decreasing mA by one-half. d. Decreasing mA by three-quarters.
Item with all relevant information in the stem	By what fraction would mA need to be decreased to lower the radiation dose by one-half? a. One-quarter. b. One-third. c. One-half. d. Three-quarters.

Figure 2.   Examples of incomplete and complete stems are given. The stem should include all relevant information and avoid repetition in the options. In the second example, the test item has been revised so that all relevant information is in the stem.

Guidelines for Writing MCQs

Several authors have outlined the elements of good MCQs (1,9,10,13,15). The National Board of Medical Examiners has published on its Web site a manual on constructing written test questions for the basic and clinical sciences; the manual reflects what the authors have learned in developing test items and tests over the past 20 years (16). Published guidelines should be viewed as best-practice rules and not absolute rules. In some circumstances, it may be appropriate to deviate from the guidelines. However, such cases should be justified and occur infrequently.

Terms are applied to the different components of MCQs. The *item* is the entire test question, which consists of a stem and several options. The *stem* is the question, statement, or lead-in to the possible answers. The possible answers are called *options, alternatives,* or *choices.* The correct option is called the *keyed response.* The incorrect options are called *distractors* or *foils.*

Writing Stems

The stem is usually composed first and is best written as a complete sentence or question. Direct questions (eg, Which of the following characteristics is an imaging feature of benign pulmonary nodules?) are clearer than sentence completions (eg, Benign pulmonary nodules. . .). Research has shown that the use of incomplete stems lowers the students' correct response rate by 10%–15% (17). A stem can incorporate maps, diagrams, graphs, or radiologic images, but these should be accompanied by a complete statement. Ideally, the item should be answerable without all of the options being read.

The stem should contain all relevant information and as much of the item as possible. If a phrase is stated in the stem, it should not be repeated in the options. Figure 2 illustrates how one test item was revised so that all relevant information is in the stem and thus avoiding the need to repeat a phrase in each option.

The stem should include only the necessary information and be kept as short as possible. It should not be used as an opportunity to teach, nor it should contain statements that are informative but not needed for the examinee to select the correct option. Stems should not be tricky or misleading, such that they might deceive the examinee into answering incorrectly. The level of reading difficulty should be kept low by using simple language so that the stem is not a test of the examinee's reading ability. As a general guide, students can complete between one and two MCQs per minute (18,19). Test items that require significantly more time to be completed should be closely examined as to whether they are unnecessarily verbose or confusing.

The stem is generally longer when application of knowledge is being tested as opposed to the recall of an isolated fact. Use of patient vignettes is a good way to test application of knowledge. Clinical vignettes can provide the basis for the question, beginning with the presenting problem of a patient; they may include the history (duration of signs and symptoms), physical findings, results of diagnostic studies, initial treatment, or subsequent findings. Vignettes do not need to be long to be effective. They should avoid verbosity, extraneous material, and "red herrings." In a study that compared the use of no vignettes, short vignettes, and long vignettes in MCQs designed to require increasing levels of interpretation, analysis, and synthesis (5), test items were shown to be more difficult as patient findings were presented in a less interpreted form. However, the differences in discrimination were not statistically significant. Regardless of these psychometric re-

Teaching Point   Ideally, the item should be answerable without all of the options being read.

Item measuring recall	Which of the following conditions manifests as chronic (longer than 3 months) airspace disease on a chest radiograph? a. Streptococcal pneumonia. b. Adult respiratory distress syndrome. c. Pulmonary edema. d. Pulmonary alveolar proteinosis.
Item with a vignette measuring application of knowledge	A 30-year-old-man presented with a 4-month history of dyspnea, low-grade fever, cough, and fatigue. Given the following chest radiograph, what is the most likely diagnosis? a. Adult respiratory distress syndrome. b. Pulmonary edema. c. Streptococcal pneumonia. d. Pulmonary alveolar proteinosis.

Figure 3.   Examples of items that test recall and application of knowledge.

Negatively worded stem	Which of the following findings is NOT a characteristic CT appearance of small airway disease?
Positively worded stem	Which of the following findings best allows small airway disease to be distinguished from interstitial lung disease on chest CT scans?

Figure 4.   Examples of negatively and positively worded stems.

Teaching Point

sults, items that incorporate vignettes are generally thought to be more appropriate because they test application of knowledge and thus improve the content validity of the examination (5). Figure 3 illustrates examples of items testing recall and application of knowledge.

The stem should be stated so that only one of the options can be substantiated and that option should be indisputably correct. It is wise to document (for later recall) the source of its validity. If the correct option provided is not the only possible response, the stem should include the words *of the following*. When more than one option has some element of truth or accuracy but the keyed response is the best, the stem should ask the student to select the *best* answer rather than the correct answer.

Questions should generally be structured to ask for the correct answer and not a "wrong" answer. Negatively posed questions are recognizable by phrases such as "which of the following is not true" or "all of the following except." Negative questions tend to be less effective and more difficult for the examinee to understand (9). Negative stems may be appropriate in some instances, but they should be used selectively. When negative stems are used, the negative term (eg, *not*) should be underlined, capitalized, or italicized to make sure that it is noticed. Figure 4 illustrates examples of negatively and positively posed questions.

Absolute terms, such as *always, never, all,* or *none* should not be used in the stem or distractors. Savvy examinees know that few ideas or situations are absolute or universally true (20). The terms *may, could,* and *can* are cues for the correct answer, as testwise examinees will know that almost anything is possible. Imprecise terms such as *seldom, rarely, occasionally, sometimes, few,* and *many* are not uniformly understood and should be avoided. In a study conducted at the National Board of Medical Examiners (5), 60 people who wrote questions for various medical specialty examinations were asked to review a list of terms used in MCQs to express frequency of occurrence and to indicate the percentage of time reflected by each term. The mean value plus or minus one standard deviation exceeded 50 percentage points for more than half of the phrases. For example, on average, the item writers believed the term *frequently* indicated 70% of the time; half believed it meant between 45% and 75% of the time; actual responses ranged from 20% to 80%. Of particular note is that values for *frequently* overlapped with values for *rarely*. Use of absolute numbers (eg, "In less than 15% of the population") is better than use of imprecise terms such as *rarely*.

Writing Options

The best number of options is three to five. Research has shown that three-option items are as effective as questions with four choices (21). Constructing questions with more than five options is burdensome and often leads to faulty options while increasing the reading demands of the student. Furthermore, there is no hard and fast rule that the number of options needs to be uniform (18). In one examination, some items may have four options and some may have five.

Item with a keyed response longer than the distractors	An otherwise healthy 28-year-old woman presented with a 2-day history of cough, fever, shortness of breath, and the following chest radiograph. What is the most likely diagnosis? a. Tuberculosis. b. Community-acquired streptococcal pneumonia. c. Varicella pneumonia. d. Blastomycosis.
Item with options of similar length	An otherwise healthy 28-year-old woman presented with a 2-day history of cough, fever, shortness of breath, and the following chest radiograph. What is the most likely diagnosis? a. Tuberculosis. b. Streptococcal pneumonia. c. Varicella pneumonia. d. Blastomycosis.

Figure 5. Examples of items with options of unequal and similar lengths.

Option C does not follow grammatically from the stem	A chest radiographic finding of left upper lobe collapse is _____. a. Hyperlucency of the upper and lower left hemithorax. b. Elevation of the left diaphragm. c. When the mediastinum shifts to the right. d. Posterior displacement of the minor fissure.
All options follow grammatically from the stem	A chest radiographic finding of left upper lobe collapse is _____. a. Hyperlucency of the upper and lower left hemithorax. b. Elevation of the left diaphragm. c. Mediastinal shift to the right. d. Posterior displacement of the minor fissure.

Figure 6. Examples of an item with an ungrammatical option and an item with all grammatical options.

Teaching Point

The most challenging aspect of creating MCQs is designing plausible distractors. The ability of an item to discriminate (ie, separate those who know it from those who don't) is founded in the quality and attractiveness of the distractors. The best distractors are *(a)* statements that are accurate but do not fully meet the requirements of the problem and *(b)* incorrect statements that seem right to the examinee (20). Each incorrect option should be plausible but clearly incorrect. Implausible, trivial, or nonsensical distractors should not be used. Ideal distractors represent errors commonly made by examinees. Distractors are often conceived by asking questions such as "what do people usually confuse this entity with," "what is a common error in interpretation of this finding," or "what are the common misconceptions in this area?"

Distractors should be related or somehow linked to each other. That is, all options should fall into the same category as the correct answer; they should either be diagnoses, tests, treatments, prognoses, or disposition alternatives. For example, all options might be a type of pneumonia or radiation dose.

The distractors should appear as similar as possible to the correct answer in terms of grammar, length, and complexity. There is a common tendency to make the correct answer substantially longer than the distractors (Fig 5).

The distractors should not stand out because of their phrasing. Grammatical cues (ie, when one or more options don't follow grammatically from the stem) can lead the examinee to the correct option (Fig 6). For example, if the stem is in past tense, all of the options should be in past tense. If the tense calls for a plural answer, all of the options should be plural. Stem and options should have subject-verb agreement. Because an item writer tends to pay more attention to the correct option than to the distractors, grammatical errors are more likely to occur in the distractors. This sort of error in test item writing is usually not an issue when the stem is written as a question.

Options should not include the phrases *none of the above* or *all of the above. None of the above* is problematic in items in which judgment is involved and in which the options are not absolutely true or false. If the correct response is intended to be one of the other listed options, knowledgeable examinees can be faced with a dilemma because they have to decide between a very detailed perfect option and the one that is intended as correct. Examinees can often construct an option

Item with overlapping options	What is the average effective radiation dose from chest CT? a. 1–8 mSv b. 8–16 mSv c. 16–24 mSv d. 24–32 mSv
Item without overlapping options	What is the average effective radiation dose from chest CT? a. 1–7 mSv b. 8–15 mSv c. 16–23 mSv d. 24–32 mSv

Figure 7.   Examples of items with and without overlapping options.

that is more correct than the one intended to be correct. Use of *none of the above* turns the item into a true-false item; each option must be evaluated as more or less true than the universe of unlisted options (16). *None of the above* only informs about what the examinee knows is not correct and not what is correct. For items with *all of the above* as a choice, the examinee only needs to recognize that two of the options are correct for *all of the above* to be the correct option.

Eponyms, acronyms, or abbreviations without some qualification after each term should be avoided. Examinees may be unfamiliar with such terms, or the terms may have more than one meaning. In such cases, the item becomes a test of whether the examinee understands the meaning of a term, or the item is faulty because a term can be interpreted in more than one way.

Options should not include material that is potentially offensive or unfair to selected groups of examinees. Therefore, references to gender or race should be made only when necessary and clinically appropriate.

Options should be placed in logical order, if there is one. For example, if the answer is a number, the options should begin with the smallest value and proceed to the largest (it is also acceptable to begin with the largest value and proceed to the smallest). If the options are dates, they should be listed in chronologic order. If the options are ranges of values, the choices should be independent and not overlap with each other (Fig 7).

Options in one item should not reveal information that allows the examinee to automatically know the correct answer to another item. This error in writing MCQs is referred to as "cueing," when an option in one item provides a hint to the answer for another item. It is also important to avoid "hinging," in which questions require that students know the answer to one item to be able to answer another item. Items must be independent of one another.

The position of the keyed response should vary among the A, B, C, and D positions. Research shows that the B or C position is overused for the correct option (21). Testwise examinees, familiar with this tendency, will choose option B or C to increase their likelihood of answering a question right when they don't know the correct answer and are forced to guess.

Item Analysis

Items that attempt to assess critically important topics cannot do so unless they are well-structured. Flaws in test questions that benefit the testwise examinee (eg, grammatical cues, use of terms such as *always* or *never,* and the correct answer being longer than the other options), and items with irrelevant difficulty (eg, long or complicated options, inconsistently stated numerical data, use of vague terms such as *rarely* or *usually,* use of *none of the above,* and tricky or unnecessarily complicated stems) must be avoided for MCQs to generate valid scores.

Several item-writing principles have been investigated for their effects on test psychometric indices (4). Most studies evaluate the effect of a single flaw in test items, such as negative stems (6) and the *none of the above* option (22). Downing (22) evaluated the validity of a classroom achievement test in medical education that contained flawed test items. Eleven (33%) of the 33 items were classified as flawed (unfocused item stems, use of *none of the above* and *all of the above*, and negative stem). He found that flawed items caused nearly one-quarter more students to fail than unflawed items. The increased test and item difficulty associated with the use of flawed items is an example of construct-irrelevant variance, because poorly crafted test questions add artificial difficulty to the test scores. This variance interferes with the accurate and meaningful interpretation of test scores and negatively affects students' passing rates, particularly for passing scores at or just above the mean of the test score distribution.

Authors of MCQs should review their test items for accuracy and appropriate formatting. However, just as with any editorial work, internal review may not reveal all errors. It can be very beneficial to have a colleague read and respond to the MCQs and offer feedback. Many medical schools have offices of medical education that can analyze the quality of test items for faculty. As MCQs become more widely used in the MOC process, organizations that provide continuing

medical education activities and SAMS should consider providing professional assistance with test item writing and item analysis. Figure 8 provides a list of guidelines for writing effective MCQs that can be referenced when proofing test items.

MCQs can be evaluated according to their reliability, validity, and resource intensiveness (23,24). *Reliability* provides a measure of an item's generalizability. Items in a test represent a small sample of all the possible MCQs that could be asked, and the test score should be indicative of the score of the same student on any other set of relevant items. *Validity* refers to the extent that a test measures what it claims to measure. *Resource intensiveness* is determined by the costs of constructing and grading items. MCQs are relatively easy to grade, especially with computer assistance, but they are difficult and time-consuming to construct.

Item analyses provide a numerical assessment of item difficulty and item discrimination. *Item difficulty* is determined from the percentage of students who answered each item correctly, with the goal being to construct a test that contains only a few items that more than 90% or less than 30% of students answer correctly (20). Optimally, difficult items are those that about 50%–75% of the students answer correctly. Items are considered low to moderately difficult if between 70% and 85% of the students select the correct response.

Item discrimination refers to the percentage difference in correct responses between two groups of students (generally referring to students in the top 25% and the lower 25%). The discrimination ratio for an item will fall between −1.0 and +1.0. The closer the ratio is to +1.0, the more effectively that item distinguishes students who know the material (the top group) from those who don't (the bottom group). Ideally, each item will have a ratio of at least +.5 (20). An item with a discrimination of 60% or greater is considered a very good item, whereas a discrimination of less than 19% indicates a low discrimination item that needs to be revised (15). An item with a negative index of discrimination indicates that the poor students answer correctly more often than do the good students, and such items should be avoided.

Test Items
—Relate directly to instructional objectives
—Test at the same level of learning as the objectives are designed to assess
—Reflect different levels of learning (recall, comprehension and application, problem solving)

Stems
—Provide a complete statement
—Include only relevant information
—Contain as much of the test item as possible
—Have a length as short as possible, commensurate with the level of learning being tested
—Ask for the correct, not the "wrong" answer
—Avoid absolute terms, such as *always, never, all,* or *none*
—Avoid imprecise terms, such as *seldom, rarely, occasionally, sometimes, few,* or *many*
—Avoid cues, such as *may, could,* or *can*
—Define eponyms, acronyms, or abbreviations when used

Options
—Follow grammatically from the stem
—Relate to each other in kind (eg, all diagnoses, all tests, all treatments)
—Are similar in grammar, length, and complexity, with distractors being plausible but clearly incorrect
—Avoid *none of the above* and *all of the above*
—Follow a logical order (eg, numeric, chronologic)
—Are independent and do not overlap
—Vary position of the correct answer

Figure 8.   Guidelines for writing effective MCQs.

Figure 8 provides a list of guidelines for test item writing and for writing effective stems and options.

Teaching Point

Summary

As the demand for continuing medical education materials and SAMs increases, so does the need for individuals skilled in item-writing. Radiologists, typically not trained in item-writing, will be one group of individuals called on to develop these materials. Radiologists are generally not

RadioGraphics

familiar with how to write measurable educational objectives and MCQs that match those objectives in terms of the level of learning involved. Beyond that, effective item construction requires knowledge of established item-writing principles. Figure 8 provides a list of guidelines for test item writing and for writing effective stems and options. This list can be referenced by radiologists who are writing MCQs for students at all levels (ie, medical students, residents, and practicing radiologists). It is important for test developers to be skilled in effective test item writing to ensure that the materials used to evaluate learners are valid assessments of a learner's knowledge. Measurement of a learner's knowledge is an important step in the educational process that should be afforded the same attention given to the development and implementation of curricula. The results of measurements of learning are used in establishing future learning goals, which completes the continuous cycle of learning.

References

1. Farley JK. The multiple choice test: writing the questions. Nurse Educ 1989;14:10–12, 39.
2. Kemp JE, Morrison GR, Ross SM. Developing evaluation instruments. In: Designing effective instruction. New York, NY: MacMillan College Publishing, 1994; 180–213.
3. Gronlund NE. Assessment of student achievement. Boston, Mass: Allyn & Bacon, 1998.
4. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines. Appl Meas Educ 2002;15:309–333.
5. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. Philadelphia, Pa: National Board of Medical Examiners, 1998.
6. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew H. The quality of in-house medical school examinations. Acad Med 2002;77:156–161.
7. American Board of Radiology. Diagnostic radiology MOC requirements. Available at: *http://www.theabr.org/DR_MOC_Req.htm*. Accessed January 18, 2006.
8. Van Hoozer H. The teaching process: theory and practice in nursing. Norwalk, Conn: Appleton-Century-Crofts, 1987.
9. Braddom CL. A brief guide to writing better test questions. Am J Phys Med Rehabil 1997;76:514–516.
10. Cox KR, Bandaranayake R. How to write good multiple choice questions. Med J Aust 1978;2:553–554.
11. Bloom BS, ed. Taxonomy of educational objectives. Vol I: Cognitive domain. New York, NY: McKay, 1956.
12. Fuhrmann BS, Grasha AF. A practical handbook for college teachers. Boston: Little, Brown, 1983.
13. Schultheis NM. Writing cognitive educational objectives and multiple-choice test questions. Am J Health Syst Pharm 1998;55:2397–2401.
14. Newble D, Cannon R. Curriculum planning. In: A handbook for teachers in universities and colleges: a guide to improving teaching methods. 3rd ed. London, England: Kogan Page, 1995.
15. Vydareny KH, Blane CE, Calhoun JG. Guidelines for writing multiple-choice questions in radiology courses. Invest Radiol 1986;21:871–876.
16. The National Board of Medical Examiners. Constructing written test questions for the basic and clinical sciences. Available at: *http://www.nbme.org/about/itemwriting.asp*. Accessed May 25, 2005.
17. Kent TH, Jones JJ, Schmeiser CB. Some rules and guidelines for writing multiple choice test items. Iowa City, Iowa: University of Iowa College of Medicine and American College Testing Program, 1974.
18. Srinivasa DK, Adkoli BV. Multiple choice questions: how to construct and how to evaluate? Indian J Pediatr 1989;56:69–74.
19. Lowman J. Mastering the techniques of teaching. San Francisco, Calif: Jossey-Bass, 1984.
20. Davis BG. Multiple-choice and matching tests. In: Tools for teaching. San Francisco, Calif: Jossey-Bass, 1993; 262–271.
21. McKeachie WJ. Teaching tips. 8th ed. Lexington, Mass: Heath, 1986.
22. Downing SM, Baranowski RA, Grosso LJ, Norcini JJ. Item type and cognitive ability measured: the validity evidence for multiple true-false items in medical specialty certification. Appl Meas Educ 1995;8:189–199.
23. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004;38:974–979.
24. van der Vleuten CP. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract 1996;1:41–67.

Teaching Points for Writing Multiple-Choice Questions for Continuing Medical Education Activities and Self-assessment Modules

Jannette Collins, MD, MEd, et al

Page 545
Thoughtfully written objectives are critical to the construction of appropriate test questions and in ensuring adequate assessment of intended learner competence.

Page 546
Ideally, the item should be answerable without all of the options being read.

Page 547
The stem should be stated so that only one of the options can be substantiated and that option should be indisputably correct.

Page 548
The best distractors are *(a)* statements that are accurate but do not fully meet the requirements of the problem and *(b)* incorrect statements that seem right to the examinee (20). Each incorrect option should be plausible but clearly incorrect.

Page 550
Figure 8 provides a list of guidelines for test item writing and for writing effective stems and options.

Test Items
—Relate directly to instructional objectives
—Test at the same level of learning as the objectives are designed to assess
—Reflect different levels of learning (recall, comprehension and application, problem solving)

Stems
—Provide a complete statement
—Include only relevant information
—Contain as much of the test item as possible
—Have a length as short as possible, commensurate with the level of learning being tested
—Ask for the correct, not the "wrong" answer
—Avoid absolute terms, such as *always*, *never*, *all*, or *none*
—Avoid imprecise terms, such as *seldom*, *rarely*, *occasionally*, *sometimes*, *few*, or *many*
—Avoid cues, such as *may*, *could*, or *can*
—Define eponyms, acronyms, or abbreviations when used

Options
—Follow grammatically from the stem
—Relate to each other in kind (eg, all diagnoses, all tests, all treatments)
—Are similar in grammar, length, and complexity, with distractors being plausible but clearly incorrect
—Avoid *none of the above* and *all of the above*
—Follow a logical order (eg, numeric, chronologic)
—Are independent and do not overlap
—Vary position of the correct answer