Question-Writing Guidelines

The multiple-choice question (MCQ) typically consists of text that provides the information required to present a problem (the "stem"), the question itself ("lead line"), and, finally, by a list of options (one correct answer and "distractors") from which the examinee chooses a response. The Board strongly encourages development of patient-based questions; in other words, questions relating to clinical scenarios. A good MCQ contains medical content presented in such a way that the candidate who possesses the required knowledge will be able to answer the question correctly; the candidate who does not possess this knowledge will be unable to do so. Correct answers must be absolutely correct. Incorrect answers should look correct to the less knowledgeable candidate. Questions should not be tricky or overly difficult; rather, they should focus on assessing the examinee's ability to provide excellent care to patients. Consider asking yourself about each question, "Is this material that a general internist or subspecialist in practice must recall daily, weekly, monthly once a year, or perhaps once in a lifetime?" Less frequently used information still is important if it deals with potentially life- or function-threatening conditions or important management decisions.

All ABIM examinations contain only A-type (single-best-answer) questions. The use of only A-types (1) simplifies and increases the efficiency of examination development, (2) simplifies examination administration and scoring. (3) makes candidate scores and feedback more meaningful and understandable, and (4) increases the proportion of questions assessing synthesis and clinical judgment, which enhances examination relevance and validity.

A. Before Writing:

1. Be sure to allow sufficient time for question writing. Many authors estimate that each question requires at least one hour of work.

2. Select and write down the **general content area** (such as recognition of congestive heart failure or evaluation of upper GI bleeding). Then write down the **testing point** (such as recognition of diastolic left ventricular failure or indications for upper GI endoscopy in the setting of upper GI hemorrhage).

   The MCQ should have one and only one testing point; testing multiple points confound the measurement information obtained about the examinee. For example, are you assessing the examinee's ability to make diagnostic inferences from the data given or the examinee's knowledge of appropriate diagnostic or therapeutic approaches to a particular problem?

   A clear testing point will help you write the question. Although it is not part of the question as presented to the examinee, a sentence or two about the question's testing point or its rationale should be determined by the question writer before the question is written.

3. Keep in mind the level of knowledge required to answer the question. Is it appropriate for the candidate? A good question is relevant and neither too hard nor too easy.

4. Think about the **cognitive ability** you wish to test that fits the testing objective. MCQs can assess the cognitive abilities of recall knowledge, synthesis, and judgment.

Recall knowledge questions require only the recall of facts; for example:

Which of the following ***malignant conditions is most likely to be cured if the patient remains*** *disease free for 30 months?*

(A) Acute myelogenous leukemia
(B) Carcinoma of the breast
(C) Ovarian carcinoma
* *(D) Embryonal cell carcinoma of the testis*
(E) Small cell carcinoma of the lung

RATIONALE: In this example, the candidate's knowledge of the prognosis of several treatable cancers is being tested. The candidate must select the malignancy that has the highest probability of cure from among the options, each of which has a reasonable response rate to treatment. Thus, the ability being tested is recall knowledge.

Synthesis questions require the integration and interpretation of facts to reach a conclusion; for example:

A 72-year-old woman has had increasing fatigue and limb weakness for one year. One month ago she was treated for pneumonia, and since that time the weakness has become much worse. She cannot rise from a chair or lift her arms to comb her hair.

Physical examination reveals weakness and hypotonia of all four limbs, with more weakness proximally than distally. There is weakness of neck flexors and extensors. No fasciculations or atrophy is noted. The muscles are not tender, and reflexes are preserved. Babinski's sign is absent. Sensory examination is normal.

Laboratory studies:

Leukocyte count	*4500/1; normal differential*
Erythrocyte sedimentation rate	*77 mm/hr*
Serum creatine kinase:	
Total	*3200 U/L*
MB isoenzymes	*21*

Which of the following is the most likely diagnosis?

(A) Amyotrophic lateral sclerosis
(B) Polyneuropathy
(C) Limb-girdle syndrome
* *(D) Polymyositis*
(E) Polymyalgia rheumatica

RATIONALE: In this example, the candidate must consider a constellation of symptoms and laboratory findings and then formulate a differential diagnosis. A list of diseases with similar presentations is given; the candidate must select the most likely diagnosis based on his or her knowledge of these entities (prevalence, clinical features). Thus, the ability being tested is synthesis.

Judgment questions require knowledge, interpretation, synthesis, and then the application of judgment to take the appropriate action; for example:

You are asked to see a 73-year-old woman who has had epigastric pain for several months. The pain is relieved by sucralfate. Her illness has not interfered with her activities.

Findings of *physical examination are normal. Upper gastrointestinal series with small bowel follow-through, obtained by the referring physician, was normal. Serum gastrin was 789 pg/mL four weeks ago and 830 pg/mL two weeks ago.*

Which of the following tests should you order next?

(A) Serum calcium level
* *(B) Measurement of gastric pH*
(C) Secretin stimulation test
(D) Upper gastrointestinal endoscopy
(E) Computed tomography of the abdomen

RATIONALE: In this example, the candidate first must recognize that the most likely condition causing this patient's symptoms is atrophic gastritis. Then the decision must be made about which test to order. Each test listed might be considered as part of the evaluation of a patient with abdominal pain and elevated gastrin; however, measurement of gastric pH would provide the necessary confirmatory information at the lowest cost and morbidity. To arrive at a correct decision, the candidate must consider issues of test sensitivity, specificity, and cost. Thus, the ability being tested is clinical judgment.

Patient-based questions that test the ability to synthesize information or demonstrate clinical judgment generally are considered more relevant than questions testing only recall knowledge. However, recall knowledge questions assess ability more directly than synthesis or judgment questions. With a recall knowledge question, the examinee either knows the requisite information or does not; with synthesis and judgment questions, the test developer does not know how the examinee arrives at a particular answer. The recall knowledge format is recommended for initially testing the knowledge base in a new or emerging discipline, because it allows test developers to accurately gauge the extent to which the discipline has been learned by the test population; in other words, it provides a baseline measure.

B. Building the Stem:

1. The "stem" is the set-up, or scenario, that leads to the question being asked. **The information in the stem should be complete, concise, clear, and unambiguous.** It should contain only the information needed to answer the question, and extraneous details should be avoided. In an exam, a patient scenario is not a "real case"; rather, it is **a convenient fiction** that sets up the testing point.

   Be sure to include the following in all patient-based questions:

   (a) Gender
   (b) Age
   (c) Site of care

   Information about race/ethnic origin and occupation should be included only if it is relevant to the testing point and cannot be answered correctly without it.

2. **Avoid tricks** to mislead examinees away from the correct answer. Often unintentional ambiguities can distract an examinee, but sometimes test developers deliberately place obstacles in the examinee's way that are not part of the testing point but can reasonably be interpreted as significant when in fact they are not (so-called red herrings). This practice is unfair to examinees and interferes with interpretation of examination results.

3. **Avoid ambiguous or indefinite terms of degree or amount,** such as *rarely, commonly, frequently, generally, sometimes,* and *usually.* These are not interpreted in the same way by all readers. Adjectives such as *young, middle-aged, older,* and *obese* are also subject to interpretation, so quantitative terms should be used (50 years old, and BMI of 31). Finally, avoid constructions that may appear pejorative, such as the use of *complains* and *denies* in the patient history.

4. **Avoid jargon** or other language that may not be known by all examinees.

5. **Avoid unnecessary ancillary material,** such as an illustration, that will consume testing time if it is not part of the testing objective. Pictures should be used only when they must be interpreted by the examinee to reach the correct answer. Illustrations or other non-textual stimuli should always demonstrate clear-cut, non-subtle findings because readers in test-taking situations tend to over interpret what they see in an illustration.

C. Posing the Question and Identifying the Task:

l. **Focus on the cognitive task** as you begin to write a question line (or "lead line"). A question should pose one clear task, such as diagnosis, or treatment, or a definition. If the stem describes a patient with many problems, the question should be focused on the problem related to the testing objective.

The question task is posed clearly if the examinee can cover up the response options and correctly deduce what some of them are. For example, in the question *Which* of *the following is the most likely diagnosis?,* the examinee could deduce that the responses will be the diagnoses that would be considered in the case presented. But in the question *Which* of *the following statements is correct about depression?,* the examinee has no clear idea what the response options will address because no specific task is posed in the question.

When more than one point is addressed, such as in a statement-based question, the assessment information is confounded because precision is lost. In addition, the examinee can easily be confused by "irrelevant difficulty"; in this case, difficulty related to question format rather than question content. (See further discussion at D.3., later.)

Here is a list of appropriate question tasks for ABIM questions:

(a) Diagnostic inference/Differential diagnosis:
Which of the following is the most likely diagnosis?
Which of the following best explains this patient's current symptoms?

(b) Clinical features:
The clinical manifestations of [disease named] *include which of the following?*
Which of the following is characteristic of this patient's illness?

(c) Diagnostic testing:
Which of the following will document the source of this patient's symptoms?
Which of the following laboratory studies should you order next?

(d) Natural history/Epidemiology:
This patient is at increased risk for the development of which of the following?
Which of the following best predicts the development of [disease named] *in a patient who has* [condition named]?
A statistically significant correlation between a history of [feature named] *and the population prevalence of* [disease named] *exists for which of the following?*

(e) Treatment:
Which of the following is most likely to correct this patient's problem?
Which of the following drugs [or therapeutic interventions] *should you order?*

(f) Management decision:
Which of the following should you do next [or now]? *Which of the following is the best management plan?*

(g) Pathophysiology/Basic science:
Gram stain of the causative organism is most likely to reveal which of the following?
Which of the following is the best explanation for this patient's poor response to therapy?
The biopsy specimen shown is consistent with which of the following conditions?

(h) Interpretation of literature/Statistical methods:
Which of the following statements best describes the findings of these researchers?
Which of the following is the best interpretation of these data?

2. **Focus the lead-in question** on your testing point:

a. Avoid using the phrase *associated with.* More specific language should be used to describe a relationship.

b. Avoid "negative" questions. The use of *Which of the following is NOT true* or *Which of the following is LEAST likely* is unacceptable, because this format requires the examinee to switch from positive to negative thinking. An example follows:

A 72 year-old black man who has insulin-dependent diabetes mellitus has had a chronic, draining ulcer of the left foot for six weeks. The wound is purulent and foul-smelling. The patient is afebrile; leukocyte count is 11,600/.L. Radiographs of the foot show osteomyelitis.

Which of the following would be LEAST appropriate as initial therapy?

* *(A) Nafcillin, intravenously*
(B) Cefoxitin, intravenously
(C) Cefotaxime, intravenously
(D) Cefazolin and metronidazole, intravenously
(E) Clindamycin and tobramycin, intravenously

CRITIQUE: In this example, the candidate is asked to use "negative" reasoning to identify the LEAST appropriate option. He or she first must recognize that a patient with diabetes and osteomyelitis requires anaerobic, broad-spectrum antibiotic coverage, and then must identify nafcillin as the least appropriate choice among the options given. This backward logic forces the candidate to switch tracks in thinking, and it fails to duplicate clinical reasoning that would focus on which drug to prescribe.

D. Developing a List of Options:

1. Make sure there is one clearly best answer. There must be one response option that the content experts agree is clearly the best answer for the situation presented. Questions in which the best answer is a matter of disagreement or controversy among experts should not be used because they cannot be scored fairly; these questions do not make it onto Board exams. Questions for which there may be multiple answers that are equally correct cannot stand as a traditional MCQ unless only one of the equally correct options is included and is keyed as the correct answer; otherwise, it is difficult for test developers to devise a scoring scheme that weights acceptable answers defensibly; and the examinee is put in a bind by having to contradict the exam instructions to pick the single best option as the answer.

2. **Add realistic, plausible distractors.** Distractors may be partially correct, but not the best answer among those listed. The distractors should reflect the realistic choices that could be considered for the situation posed; they may reflect common misconceptions, outdated beliefs, or commonly confused ideas. If you are unable to come up with adequate distractors, then the question will not work, regardless of how realistic the scenario or how important the content.

   Guessing is not a significant contributor to test scores at this level of examination, so including implausible, trivial, or nonsense distractors will only weaken the question. Three or four good options are better than five options that include one or more nonsense distractors.

3. **Avoid combining right and wrong in the same distractor.** This puts examinees in a bind. For example, in a treatment question in which there are two equally good drugs for a given condition, making one of the options wrong by specifying the wrong dosage.

4. **Avoid irrelevant clues to the correct answer.** Here are some ways to avoid giving the candidate a clue to the answer of your question:

   Do not create implausible, obvious, or nonsense distractors. For example, in a test of ethics, do not include an obviously unethical choice as a response option.

   Do not make the correct answer too attractive to resist. By including obviously appropriate management recommendations (such as to stop smoking and start exercising) only in the correct answer and not in the distractors, the answer is given away.

   Do not make the correct answer substantially longer or more detailed than the distractors. All the response options should be homogeneous and similar in grammatical construction, length, and complexity. It is a common belief among test-savvy candidates that the longest option is the correct answer.

   Do not create two distractors that illustrate the same wrong way of thinking. When two distractors are very similar in meaning or intent, this is a clue that neither of them can be the correct answer.

5. **Do not use non-homogeneous options.** These questions ask the candidate to choose between apples and oranges; that is, to select the best option from a list of options addressing different points. For example:

> *An 87 year-old white man who has severe longstanding Alzheimer's disease now resides in a nursing home because of his dementia and total dependence in all activities* of *daily living. He has a long history of recurring constipation and laxative use. Several episodes offecal impaction have been relieved manually. During the past year, despite a treatment program that included dietary fiber, increased* f *fluid intake, toileting schedules, and enemas, he has required three urgent hospital visits for abdominal problems that included constipation, fecal impaction, and abdominal distention. On each of these occasions, radiographs have shown massive colonic dilation.*

Which of *the following statements regarding this patient's gastrointestinal syndrome is true?*

(A) *The first symptom is usually abdominal pain*
(B) *The use* of *fiber supplements is essential in treatment*
(C) *Effective treatment requires cleansing tap water enemas every three to four days*
(D) *It may be exacerbated by nonsteroidal anti-inflammatory drugs (NSAIDs)*

Every question should address a single content point. When more than one point is addressed, the assessment information is confounded because precision is lost. More importantly, the examinee can easily be confused by such a question because it introduces "irrelevant difficulty" (in this case, difficulty related to question format rather than question content). In your question, the essence of the problem should be in the stem, which means that the question should pose a specific, single task such as diagnosis, treatment, and so forth. One way to identify any well-constructed question is first to cover the options and then to guess what they might be from the question asked. If some of the options can be deduced, it is clear that a specific, meaningful task is being posed. In the question above, no specific task is posed; therefore it would be almost impossible to deduce any of the options.

6. **Do not use *"All of the above."*** This is an unacceptable option because it increases the chance that a less knowledgeable examinee will guess the correct answer.

7. **Do not use *"None of the above."*** This is an unacceptable option because it is an imprecise measure of the candidate's knowledge. When *None* of *the above is* the correct answer, it may be chosen by a candidate who thinks it refers to another "answer" that is not be correct. When *None of the above is* a distractor, it may mislead a candidate who is thinking of an unlisted but possible correct alternative to the correct answer.

E. **After Completing Questions**:

1. Experienced authors suggest setting questions aside for a week or more, then returning to them for a fresh review.

2. Consider adding a brief rationale that describes (1) the testing point, (2) why you picked the different options, and (3) why the indicated answer is best. Time spent writing a rationale may prove useful when the time comes to discuss your question with the committee. If the question tests a new or controversial area, a reference also may be appreciated by other committee members.