Writing Multiple-Choice Questions

Robert J. Boland, M.D.
Natalie A. Lester, M.D.
Eric Williams, M.D.

Academic psychiatrists are often asked to write multiple-choice questions. Some will serve on question-writing teams for standardized examinations, while others may be required to submit questions for course examinations or continuing medical education activities. Multiple-choice tests are popular because of their practicality in testing a wide body of knowledge and the ease of scoring responses, but criticisms are many. For example, they encourage guessing, require only recognition rather than recall of details, and evaluate only a fragmentary memorization of information rather than a deep understanding of applied knowledge. However, the well-written multiple choice question differentiates those who know the material from those who do not, and may emphasize analysis of information rather than direct recall.

Psychiatrists writing for high-stakes examinations (e.g., the United States Medical Licensing Examinations or the American Board of Psychiatry and Neurology written examinations) undergo training in question construction. Many others have little training beyond their own experience in taking such examinations. Even experienced question writers may be unfamiliar with the literature in this area, because the bulk is found in nonmedical journals. As a result, questions written for medical education and evaluation are frequently flawed (1, 2).

This article will summarize the literature on multiple-choice questions, recommend how to write proper questions, and examine the literature that informs these recommendations.

Received July 29, 2009; revised October 19, 2009; accepted October 29, 2009. Dr. Boland is affiliated with the Department of Psychiatry and Human Behavior at the Warren Alpert School of Medicine at Brown University in Providence, Rhode Island; Dr. Lester is affiliated with Psychiatry and Human Behavior at the Warren Alpert Medical School at Brown University in Providence; Dr. Williams is affiliated with the Department of Psychiatry at the University of South Carolina School of Medicine in Columbia, South Carolina. Address correspondence to Robert J. Boland, M.D., Brown University, Department of Psychiatry and Human Behavior, Residency Training/Butler Hospital, 345 Blackstone Blvd, Providence, RI 02906; robert_boland_1@brown.edu (e-mail).

Copyright © 2010 Academic Psychiatry

Standards

Although multiple-choice tests have enjoyed almost a century of popularity, research on their use has developed slowly, and question-writing guides are grounded more in experience than in evidence. Interest in testing methods grew in the late 1970s following the first conference on Measurement and Methodology in Education (3). In the late 1980s, Haladyna and Downing (4, 5) wrote several influential reviews of multiple-choice question writing. These, along with a later update (6) and several book compilations (7–9) provide the basis for most test-writing guides.

The most important research question is whether "properly constructed" questions perform better as evaluation tools than purportedly "flawed" questions. In assessing this, testing experts apply psychometric tests to assess a question's performance; these analyses are beyond the scope of this article and the usual test writer does not apply them. However, a test writer should be familiar with the fundamental properties that are considered when assessing a question's quality: discrimination and difficulty.

Discrimination refers to a question's ability to differentiate between those who know the material and those who do not. Poor discrimination is akin to a type I error in that examinees who do not know the material are able to guess correctly with a frequency similar to those who know the material. This is a problem with poorly worded questions that lead an uninformed person to the correct answer.

Difficulty refers to the likelihood that a person who knows the material will answer the question correctly. A question that is "too difficult" is akin to a type II error in that persons who actually know the material may be confused by the question and answer incorrectly. This is a

TABLE 1. Faulty and Improved Questions

Faulty Question	Improved Question
Which of the following defense mechanisms is defined correctly? (A) Splitting: the emergence of multiple personalities during therapy (B) Repression: reverting to childlike behaviors (C) Displacement: perceiving that another individual has a trait that one finds unacceptable in oneself (D) Reaction formation: a strong negative reaction toward the therapist will be enacted by refusal to show to the next appointment (E) Projective identification: a trait of the patient is identified by the patient as belonging to the therapist, and then the therapist enacts this trait	A psychiatric resident has been seeing a patient with major depressive disorder for weekly psychotherapy sessions. The resident observes that her patient sees himself as helpless and rejects her suggestions during their therapy sessions. In turn, the resident feels helpless herself. This is an example of which of the following defense mechanisms? (A) Splitting (B) Repression (C) Displacement (D) Reaction formation (E) Projective identification

The first question has a number of flaws that will be explored in this article: it is unfocused, heterogeneous, and tests multiple concepts in an abstract manner. The second question improves on the first by focusing the question on a single concept that is apparent in the stem and applying it to a practical situation.

problem with trick questions that purposefully obscure the correct answer or add ambiguous details to confuse the examinee.

The question of whether the mechanics of question writing is important rests on whether differences in wording will result in differences in the discrimination and/or difficulty of a question. In general, research suggests that wording does matter, and when different questions measuring the same concept are compared in similar populations, certain styles of questions perform better than others (1, 10). This research has informed the current guidelines for multiple-choice question writing, such as the test-writing manual used by the National Board of Medical Examiners (11). Table 1 shows a flawed question and its improved version.

We will limit our discussion to the conventional multiple-choice format, which is most common in medical education. It consists of a *stem*, which poses the question, followed by several possible answers. One, called the *key*, is the correct answer; the incorrect alternatives are called *distracters*.

Overall Qualities of Good Questions

Good questions discriminate between those who know the material and those who do not and are appropriately difficult for the population being tested. Question writers first should consider what exactly is being assessed. A question should test important information that the examinee can reasonably be expected to know. Although this seems obvious, psychiatry question writers may find it difficult to tap core knowledge and instead resort to obscure biological facts or difficult to recall (but easy to test) statistics. Consider the following example:

> The direct and indirect costs of schizophrenia to the U.S. economy, including the cost of treatment, lost productivity, and mortality, is estimated to be closest to what annual dollar amount?
>
> (A) $50 million (B) $500 million (C) $5 billion (D) $50 billion (E) $5 trillion

Although answer D may be correct, the importance of knowing this data offhand is not clear, and most psychiatrists would have to look it up.

Questions should reflect the level of training. It may be difficult for a seasoned professional to gauge what is appropriate. In a classroom setting, writers should consider what was emphasized. Textbooks that are intended for certain levels (e.g., medical students, residents) may help clarify the appropriate level of difficulty.

Questions should be defensible. The first-time question writer will often be amazed at the number of "commonly accepted facts" that have limited or no support. For example:

> A 50-year-old man with major depressive disorder and no history of prior suicide attempts has a near-lethal overdose shortly after starting sertraline. Which of the following is the most likely reason for the suicide attempt?
>
> (A) Sertraline provided him with the energy to act on prior suicidal thoughts. (B) He most likely had no suicidal thoughts before starting sertraline. (C) He would have attempted suicide even without the addition of sertraline. (D) Without sertraline, he would have been more likely to complete suicide. (E) Without sertraline, he would have attempted suicide by more violent means.

Answer A was intended to be the correct answer; however, the mechanism by which antidepressants might cause increased suicide risk, while long speculated to be related to increased energy, has not been

TABLE 2. Unfocused and Focused Stems

Unfocused Stem	Focused Stem
Which of the following is true of fluoxetine? (A) It was approved by the Food and Drug Administration in 1977. (B) It is structurally a heterocyclic compound. (C) A common side effect is weight loss. (D) It is fully metabolized in about 48 hours. (E) It requires a 4-week "wash-out" period before starting an MAOI.	Switching antidepressant agents from fluoxetine to phenelzine requires a 4-week "wash-out" period due to increased risk of developing which of the following? (A) Stevens-Johnson syndrome (B) Serotonin syndrome (C) Neuroleptic malignant syndrome (D) Agranulocytosis (E) Malignant hyperthermia

The point of the first question is not clear from the stem. In the second question, the knowledge being sought is clear to the reader, and many examinees will be able to predict the answer without having to read the possible choices.

borne out by research (12). Moreover, the relationship between antidepressants and suicide risk remains controversial. The material being tested in this question is probably too complex and not well enough understood to be captured in a multiple-choice format. Preferably, questions should be supported by unambiguous references. An inability to find a reference should make one suspect a question's validity.

When writing questions, one should be careful about stereotyping race, gender, or other factors. For example, not all depressed patients should be women, and questions about violence should not center on African Americans. One way to guard against this is by using gender, race, and other descriptors only when they are germane to the question or when constant substitution of "the patient" makes the question awkward.

Writing Good Stems

A good stem should be focused, as Table 2 demonstrates. Questions with unfocused stems are confusing and reward recognition over recall. In the ideal situation, the test-taker can cover up the answer choices and generate the correct answer independently. In another type of unfocused stem, many irrelevant details, often called "window dressing," are included. For example:

> A 30-year-old physically healthy patient with a diagnosis of major depressive disorder is currently experiencing another episode of depression. The patient admits to missing doses two to three times a week due to a hectic home and work schedule. A request is made for a long-acting shot, "like with Haldol." On assessment, the patient appears generally healthy with stable vital signs and an unremarkable physical examination. You explain that, although there is not currently an antidepressant injection, you can offer the medication with the longest half-life. Which of the following would be prescribed to this patient?
>
> (A) Citalopram (B) Fluvoxamine (C) Paroxetine (D) Fluoxetine (E) Sertraline

The details only serve to obscure the main point of the question, and it becomes difficult to understand what information is being sought. The stem could be stated more clearly as "Which of the following SSRIs has the longest elimination half-life?"

A related flaw is the stem that asks more than one question, such as,

> Which of the following SSRIs has an active metabolite and the longest half-life?
>
> (A) Citalopram (B) Fluvoxamine (C) Paroxetine (D) Fluoxetine (E) Sertraline

The correct answer depends on two different factors, both of which must be true. This is sometimes called a

TABLE 3. Stem Formats

The Sentence-Completion Method	The Question Format
After 1 week of twice-daily haloperidol therapy, a patient complains of restlessness and an uncomfortable need to move. The patient is most likely experiencing (A) Tardive dyskinesia (B) Extrapyramidal symptoms (C) Akathisia (D) Neuroleptic malignant syndrome (E) Orthostatic hypotension	After 1 week of twice-daily haloperidol therapy, a patient complains of restlessness and an uncomfortable need to move. Which of the following side effects is the patient most likely experiencing? (A) Tardive dyskinesia (B) Extrapyramidal symptoms (C) Akathisia (D) Neuroleptic malignant syndrome (E) Orthostatic hypotension

TABLE 4. Adjusting the Phrasing of Questions

Negatively Phrased Question	Rephrased Question	Deeper Level Question
A 40-year-old woman with a history of breast cancer is taking tamoxifen. She presents for treatment of depression. Which of the following antidepressants should *not* be prescribed? (A) Desvenlafaxine (B) Amitriptyline (C) Mirtazapine (D) Paroxetine (E) Venlafaxine	A 40-year-old woman with a history of breast cancer is taking tamoxifen. She presents for treatment of depression. Which of the following antidepressants is most appropriate for this patient? (A) Sertraline (B) Fluoxetine (C) Paroxetine (D) Venlafaxine (E) Escitalopram	Paroxetine lowers the efficacy of tamoxifen by which of the following mechanisms? (A) Binding to and inactivating the tamoxifen molecule (B) Decreasing the metabolism of tamoxifen to its active form (C) Blocking the absorption of tamoxifen by the gut (D) Competing with tamoxifen at the receptor site (E) Disabling the second messenger system used by tamoxifen

double-barreled question. A wise test taker can eliminate distracters based on which SSRIs do not have an active metabolite and which have a short half-life. This is, in essence, two different questions with the same answer.

Stems are usually in one of two formats: a full sentence question or a phrase that requires sentence completion (Table 3). Some evidence suggests that the sentence-completion method is more difficult to understand, and for this reason Haladyna and Downing (5) prefer using the question format. It is usually simple to rewrite a sentence completion question as a full sentence. Both question types are currently used on medical examinations.

Posing the Question

In psychiatry, as with most medical fields, it is difficult to rule out all possible alternatives. Multiple-choice questions, with their implication that only one of the possible choices can be correct, open the possibility for the industrious examinee to search the literature for an obscure study that supports one of the distracters or to find literature that supports an alternative apart from any of the choices offered. These problems are controlled (although not eliminated) by posing questions in a variation of "Which of the following is *most* correct?" This format limits the possibilities to those being asked and allows that, although more than one answer might be supportable, only one is *best* supported.

Negatively phrased questions (e.g., "Which of these is *not* correct" or "all of the following are true *except*") are generally discouraged, and many standardized examinations prohibit them altogether. The argument is that they make the question unnecessarily difficult. The examinee must shift from the usual task of finding correct answers to finding incorrect ones.

Test writers have been particularly reluctant to give up these items because they are easier to write, requiring only one plausible distracter instead of multiple ones. Data on these items are mixed, and some find no difference in performance when positively and negatively worded questions are compared (6). However, most of this research was done with younger students, and some have found negatively worded items are more confusing as questions become more sophisticated (13). Other data suggest that negatively phrased items may overestimate one's ability (14).

Although the data are not conclusive, we argue that negative stems should be avoided because they are like trick questions at more advanced educational levels, and most standardized medical examinations do not allow negative questions. Negatively phrased questions can usually be changed to positively framed questions by choosing one of the distracters as the correct answer, changing the key to a distracter, and creating additional distracters (Table 4).

Vignettes

Medical examinations are trying to move from testing fragmentary facts to applied knowledge. Case vignettes are a common method of testing the latter. However, these can be problematic. Many novice test writers will attempt to write up actual cases only to find that the limitations of the multiple-choice format cannot incorporate the size and complexity of genuine cases. Generally, cases have to be simplified to the essential points so that only a single correct answer is plausible (Table 5).

Distracters

Writing good distracters is difficult. To distinguish those who know the correct answer from those who are guessing, one must create distracters that are both plausi-

TABLE 5. Vignettes

Complex	Streamlined
A 40-year-old pediatrician spends 50 hours per week working in an outpatient clinic and 20 hours per week working in an emergency room on nights and weekends. Although her colleagues will cosign clinic notes dictated by residents, she insists upon dictating her own notes because she feels that residents' notes have poor grammar and are not as thorough as she would like. She has never been married and explains that she does not have time for a significant relationship because of her long hours at work. She admits that she feels unfulfilled in this area of her life and avoids dating because she feels afraid of rejection. She had several relationships in the past, including two she would consider serious: one during medical school, which she broke off because she felt that the person was not ambitious enough, and one in her mid-30s. She is unaware of why the second person broke off the relationship, saying, "He just seemed to lose interest." Presently, she avoids most social gatherings because she fears others will think critically of her. She has few friends and enjoys few activities other than work. She lives by herself in a small, inexpensive apartment. Which of these following disorders is the most likely diagnosis? (A) Social phobia (B) Obsessive-compulsive disorder (C) Schizoid personality disorder (D) Narcissistic personality disorder (E) Obsessive-compulsive personality disorder	A 40-year-old pediatrician spends approximately 70 hours per week at work. She insists upon dictating her own notes, because she feels residents' notes have poor grammar and are not as thorough as she would like. She has never had a significant relationship due to her work and her tendency to find fault with others. She enjoys few activities outside of work. Which of the following is the most likely diagnosis? (A) Social phobia (B) Obsessive-compulsive disorder (C) Schizoid personality disorder (D) Narcissistic personality disorder (E) Obsessive-compulsive personality disorder
Although this vignette has the complexity one would associate with a person suffering from obsessive-compulsive personality disorder, the length and level of detail make the question unnecessarily difficult. Furthermore, such complex questions invite multiple interpretations of details, and more than one answer becomes plausible. It is preferable to streamline the case to the most essential details supporting the correct choice.	

ble and wrong. Furthermore, one must create distracters that do not inadvertently hint at the correct answer. There are many pitfalls to distracter writing, and the "good test taker" is able to take advantage of these.

To avoid these pitfalls, distracters should be independent. The stem should contain the bulk of the content; distracters should contain only the possible answer. Unfocused stems often use long and complex distracters. Such questions can both confuse those who know the material and cue those who do not. For example:

> Venlafaxine works primarily by reuptake of which of the following neurotransmitters?
>
> (A) Serotonin and histamine (B) Dopamine and norepinephrine (C) Serotonin and norepinephrine (D) Norepinephrine and acetylcholine (E) GABA and serotonin

In this question, the clever test taker will notice that norepinephrine and serotonin occur more often as choices than the other neurotransmitters and, using a strategy called *convergence*, correctly assume that the answer containing them both is more likely to be correct.

Distracters also should not overlap or contradict each other; if they do, examinees may logically exclude one or more of them. For example:

> A 10-year-old boy with attention-deficit/hyperactivity disorder and Tourette's syndrome is started on methylphenidate. Which of the following effects would methylphenidate be expected to have on the symptoms of the Tourette's syndrome?
>
> (A) Increase the frequency of tics (B) Decrease the frequency of tics (C) Reduce coprolalia (D) Induce choreiform movements (E) Prevent tic attenuation during sleep

The attentive test taker will realize that answers A and B are opposites and narrow the question down to a 50–50 guess.

Even among experts there is surprising disagreement as to the appropriate range of frequency for vague quantifiers such as usually, mostly, or rarely (15). When "none of the

above" or "all of the above" represents the correct answer, it undermines the purpose of an examination, because it is possible to correctly answer the item without having a full understanding of the material (16). This is particularly true for "all of the above," for which research on its use is extensive. A test taker need only recognize that more than one answer is correct to conclude that all must be correct. When "all of the above" is used, examinees tend to choose it; thus it has a low ability to discriminate between test takers (6).

When "none of the above" is an option, the question requires the test taker to compare the universe of alternative possibilities against the supplied options. Research has generally found that this option makes a question more difficult while not improving its discriminative ability. However, some educators feel that, when carefully used, "none of the above" can be appropriate, and some research supports this (17). Other studies (18) suggest that the phrase has reasonable difficulty and discrimination. At least one study suggests it improved discrimination by reducing the likelihood of guessing correctly (19).

Distracters are less likely to hint at the correct answer if they are similar in content, length, and grammatical construction, that is, *homogeneous*. An unfocused stem often leads to *heterogeneous* distracters that are unrelated to each other. Most well-informed test takers know that the longest potential answer is usually the correct one. When the items are grammatically different, options can be eliminated merely because they do not follow grammatically from the question being asked. The importance of distracter similarity has been empirically validated; homogenous items increase both difficulty and discrimination (20).

Conventionally, most standardized tests offer five choices: four distracters and a key. Research suggests that three distracters (hence, four possible answers) is the optimal number (21, 22) and that additional distracters increase difficulty but do not increase discrimination (23). Fewer distracters would be welcomed by test writers, who often find it difficult to create plausible but incorrect answers. In our examples we have adhered to the convention of five choices, the number typically found in standardized tests used in medical education. However, we recommend that test writers not be afraid to use fewer distracters when additional plausible ones cannot be added.

When possible, distracters should be placed in a logical order (for example, in a number series or successive stages of an illness). Otherwise they are generally distributed randomly. True random distribution protects against the tendency of many test writers to bury the correct answer in the middle of the group of distracters.

TABLE 6. Selected Qualities of Good Multiple Choice Questions

Overall, multiple-choice questions should
Test important material
Be appropriate to the level of training
Be supported by data or references
Stems should
Be focused and clear
Contain the majority of information
Lead to only one possible answer
Be positively phrased
Distracters should
Be short and to the point
Be independent
Be free of vague quantifiers such as "usually," "mostly," "rarely"
Avoid "all of the above" or "none of the above"
Be similar in content, length, and grammar

Conclusion

Multiple-choice questions are meant to distinguish those who know the material from those who do not. Well-written questions decrease the risk of discrimination errors. Unclear, confusing, or "trick" questions only mislead the examinee. Many guidelines exist, with most reflecting collective wisdom as well as evidence. Those in the forefront of this work have underscored the need for more research. Much of the existing research examines younger learners and may not be relevant to graduate medical education (24). Medical students and residents are particularly experienced at standardized testing and often able to compensate for flaws in test writing. However, in doing so, they may rely on strategies that have little to do with learning the material (25).

Our recommendations for good question writing are summarized in Table 6. These are not complete but meant to distill from the many recommendations those that will be most relevant to writers in medical education settings. There is insufficient evidence to justify creating ironclad rules. Novice test writers are expected to adhere more closely to these guidelines, whereas expert writers may find times when it is appropriate to go beyond the recommendations, but that should only be attempted when it best serves the goal of creating clear and fair examinations.

Finally, as some authors have pointed out (22), too much attention to the mechanics of test writing can be

misguided. The primary concern remains one of content, and the successful test writer should first consider what is important for trainees to know before considering how to ask it.

The authors thank Ed Michener and Susan H. Couch for supporting research for this article. At the time of submission, the authors reported no competing interests.

References

1. Tarrant M, Ware J: Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 2008; 42:198–206
2. 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
3. Baker EL, Quellmalz ES, University of California Los Angeles Center for the Study of Evaluation: Educational Testing and Evaluation: Design, Analysis, and Policy. Beverly Hills, Calif, Sage Publications, 1980
4. Haladyna TM, Downing SM: A taxonomy of multiple-choice item-writing rules. Appl Meas Educ 1989; 1:37–50
5. Haladyna TM, Downing, SM: The validity of a taxonomy of multiple-choice item-writing rules. Appl Meas Educ 1989; 1:51–78
6. Haladyna TM, Downing SM, Rodriguez MC: A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ 2002; 15:309–334
7. Haladyna TM: Developing and Validating Multiple-Choice Test Items. Hillsdale, NJ, Lawrence Erlbaum Associates, 1994
8. Downing SM, Haladyna TM: Handbook of Test Development. Mahwah, NJ, Lawrence Erlbaum Associates, 2006
9. Kehoe J: Writing Multiple-Choice Test Items. Practical Assessment, Research & Evaluation, 4(9): Office of Educational Research and Improvement (ED), Washington, DC, 1995, Report No EDO-TM-95–3
10. Downing SM: The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ 2005; 10:133–143
11. Case SM, Swanson DB: Constructing written test questions for the basic and clinical sciences. Philadelphia National Board of Medical Examiners, 2002. Available at www.nbme.org/publications/item-writing-manual.html
12. Shelton R, Lester N: Selective serotonin reuptake inhibitors and newer antidepressants, in The American Psychiatric Publishing Textbook of Mood Disorders. Edited by Stein DJK, Schatzberg AF. Washington, DC, American Psychiatric Publishing, 2005
13. Tamir P: Positive and negative multiple choice items: how different are they? Studies in Educational Evaluation 1993; 19:311–325
14. Harasym PH: Negation in stems of single-response multiple-choice items: an overestimation of student ability. Evaluation and the Health Professions 1993; 16:342–357
15. Case SM: The use of imprecise terms in examination questions: how frequent is frequently? Acad Med 1994; 69:S4–6
16. Gross LJ: Logical versus empirical guidelines for writing test items: the case of "none of the above." Evaluation and the Health Professions 1994; 17:123–126
17. Frary RB: The none-of-the-above option: an empirical study. Appl Meas Educ 1991; 4:115–124
18. Knowles SL, Welch CA: A meta-analytic review of item discrimination and difficulty in multiple-choice items using "None-of-the-above." Educ Psychol Meas 1992; 52:571–577
19. Kolstad RK, Kolstad RA: The option "none of these" improves multiple-choice test items. J Dent Educ 1991; 55:161–163
20. Ascalon ME, Meyers LS, Davis BW, et al: Distractor similarity and item-stem structure: effects on item difficulty. Appl Meas Educ 2007; 20:153–170
21. Trevisan MS: Estimating the optimum number of options per item using an incremental option paradigm. Educ Psychol Meas 1994; 54:86–91
22. Taylor AK: Violating conventional wisdom in multiple choice test construction. College Student J 2005; 39:141
23. Swanson DB, Holtzman KZ, Clauser BE, et al: Psychometric characteristics and response times for one-best-answer questions in relation to number and source of options. Acad Med 2005; 80:S93–96
24. Rachor RE, Gray GT: Must all stems be green? A study of two guidelines for writing multiple choice stems. Annual Meeting of the American Educational Research Association, New York, April 8–12, 1996
25. Triska OH: Clinicians' perceptions of medical students' reasoning on multiple choice items. Annual Meeting of the National Council on Measurement in Education, April 11, 1996, New York, p 16