SPECIAL ARTICLE

A guide to writing ABO test items

Barry S. Briss,[a] Jeryl D. English,[b] Michael L. Riolo,[c] and Peter M. Greco[d]
Boston, Mass, Houston, Tex, Grand Haven, Mich, and Philadelphia, Pa

Credentialing examinations are frequently the primary method for determining a candidate's competency before entering a profession or achieving certification. "Credentialed" implies that the certificate holder is sufficiently competent to ensure that the health, welfare, and safety of the public are protected. The testing process is an outcome assessment tool that measures areas of mandatory and essential knowledge.[1] Because credentialing examinations are only as good as the quality of their testing items, these items must accurately reflect the knowledge, skills, and abilities (KSAs) essential for entry-level competency. The primary purpose of this article is to assist in writing relevant and reliable test items.

A testing organization should precisely define both the educational objective of the examination and the information that the candidate should know. With regard to the American Board of Orthodontics (ABO) and American Dental Association, these details can be found at www.Americanboardortho.com and www.ada.org.Tests should give the examining body enough information to assess all applicable areas of knowledge[1] and have a sufficiently broad scope of questions to ensure reliability. No test should be designed until the examiners know how the results are to be used. The ABO uses the Phase II examination results to assess didactic and clinical knowledge as a part of the certification process.

When the purpose of the test has been established, an appropriate format should be chosen. The multiple-choice format is objective, versatile, and useful for testing the knowledge of a large group of people in various subjects. Therefore, the ABO chose it for the Phase II examination.

[a]Chairman, Orthodontic Department, Tufts University School of Dental Medicine, Boston, Mass.
[b]Chairman, Orthodontic Department, University of Texas Dental Branch, Houston, Tex.
[c]Adjunct professor, Orthodontic Department, School of Dentistry, University of Detroit Mercy, Detroit, Mich.
[d]Clinical associate professor, Orthodontic Department, University of Pennsylvania School of Dentistry, Philadelphia, Pa.
Reprint requests to: American Board of Orthodontics, 401 N Lindbergh Blvd #308, Saint Louis, MO 43141; e-mail, Chris@americanboardortho.com.
Am J Orthod Dentofacial Orthop 2005;128:397-401
0889-5406/$30.00

doi:10.1016/j.ajodo.2005.07.010

Objective, multiple-choice tests have the following advantages.

- They sample knowledge with maximum efficiency and reliability.
- They ask questions and elicit answers (multiple choice, true or false, fill in, matching).
- They require the candidate to read, evaluate, and respond.
- They are easy to score.
- They are efficient because they allow many items to be administered in a short time.
- They can assess important and relevant areas of knowledge.
- They can identify areas of deficiency in the candidate's knowledge.
- They can be used to identify weaknesses in program curricula.[2]

Objective tests have the following disadvantages.

- They can be poorly written, impeding a knowledgeable candidate.
- They can be biased against poor test takers.
- They can overemphasize recall and memorization.[3]

For an examination to be reliable and valid for measuring knowledge and competency, much time and effort must be devoted to developing the test. Before an item appears in an examination, hours might have been spent on its development, review, trial, and analysis. Regardless of this effort, it is still possible for the item to be flawed. By becoming familiar with the terminology and concepts of good test writing, the process will proceed more smoothly.

CONSTRUCTING ITEMS

Writing a good test item is not a science in the strictest sense, but certain conventions should be followed. A multiple-choice item consists of an introductory question or statement—the stem—followed by 3 to 5 alternative responses—the options. The stem states the central problem, suggests an idea, or asks a question. It can include graphics or describe a particular situation (a story problem) to assist the candidate in defining the problem. Following are 2 examples of simple stems.

The ABO was established in what year? (question)
The ABO was established in _____. (incomplete statement)

Options are the possible answers from which a candidate selects the correct one. The incorrect options are called foils or distracters, and the correct response is the key. Four options are generally preferred because it can become difficult to write plausible foils beyond that number. And, if only 3 options seem viable, adding an obviously incorrect fourth choice does not increase the reliability of the item. In simulated tests evaluating knowledge vs guessing in objective tests, 3 options performed similarly to 4 in terms of consistency and discrimination. The influence of test-wiseness is actually reduced with 3 options rather than 4.[4,5]

When the item is finally put together, it looks like this:

(stem) *The American Board of Orthodontics was established in what year?*
(options)
a. 1904 (distracter)
b. 1916 (distracter)
c. 1927 (distracter)
d. 1929 (key)[6]

CLASSIFYING ITEMS

Each item should relate to a predetermined set of KSAs. If the examining body wants to ensure that the test assesses appropriate aspects of a candidate's educational experience and knowledge, it should analyze the aspects that are deemed important. Educators, board-certified practitioners, and educational test consultants should work together to determine the most critical clinical-care tasks required for competency in orthodontics. The tasks determined to be relevant in the examination, as well as the corresponding KSAs resulting from the job analysis, are then incorporated into the test specifications or examination blueprint. These specifications indicate the total number of items in the examination, the major categories (tasks), the KSAs measuring each task, the number of items per category, and the cognitive or thinking level required. All items in an examination must relate directly to the predetermined test specifications so that all candidates, regardless of when they are tested, will be examined on the same concepts with the same number and type of items.

It would be easy to approach the construction of test items purely from the mechanical viewpoint. One could simply decide what a good question might be, construct it, create 3 foils and a correct answer, and call it good. However, a more complex variable should not be ignored: what aspect of learning is to be tested, or what cognitive level of thinking is required to pass the test?[7]

Learning can be categorized into several broad topics: recall, comprehension, application, in-depth knowledge, synthesis of information, evaluation, and analysis.[8] These categories should be considered when writing or reviewing test items and should be discussed as a part of the rationale for an item. For example, when asking a question that requires the candidate to use the cognitive skill of recall, the examiner might only be testing ability to memorize facts rather than ability to reason. If, on the other hand, the examiner's intent is to test ability to synthesize information, the question must be structured to elicit the response that requires that skill level.

The 3 levels of cognitive thinking generally included in credentialing examinations are recall, application, and analysis. *Recall* is the ability to remember and recognize facts, definitions, steps, and rules from previously learned material but not necessarily relating the information to anything else. Some key verbs that might be used in a recall item are choose, define, describe, identify, label, or select. *Application* refers to the ability to apply information to a new situation; to draw upon facts, principles, or steps; or to solve a problem usually in a straightforward manner. These items measure the candidate's ability to interpret or apply limited data. Some key verbs in an application item are change, make, modify, operate, prepare, produce, solve, or organize. *Analysis* is the ability to evaluate data, solve problems, fit pieces together to form a whole, and analyze the various parts and their relationship to the whole. Some key verbs in analysis items are appraise, compare, assess, or judge.

To classify the cognitive level of an item, one must ask: what is the candidate expected to do? If he or she must identify or recognize the correct answer, that item should be classified as using cognitive ability to recall information. If the candidate must classify, explain, or differentiate, the item will be described as application. If the candidate must formulate, evaluate, or judge, the item will require the cognitive ability to analyze. These levels of thinking are interrelated and build on each other. For example, to answer an analytical question, a candidate might need to recall and apply certain facts.

RULES OF WRITING ITEMS

One factor to keep in mind when writing items is that each is a sample of what a person must know to be deemed competent in the specialty. Items should not be written so that they provide clues to the correct answer, nor should they be written so poorly that they mislead the candidate into choosing the wrong answer.

Here is an item checklist:

1. Each item should measure a candidate's knowledge of a topic that is relevant to the specialty and to the protection of the public or client. Does the item measure something that a competent, entry-level candidate should know?[9]
2. Each item should test a single concept or idea.[9] Keep the purpose of the item clearly in mind and make certain that it is related to an important job component. Candidates should not be tested on trivia or knowledge that would be acquired only through experience. The item should represent current and optimal practice levels. Avoid items that measure something so new that few people are doing or using it or something so obsolete that only a few people still use it.
3. The language of the item should be appropriate for the candidate's educational or reading level. It is generally better to write an item at or below the reading level of the average candidate.[8,9] Do not make an item difficult to read; a credentialing examination is not a test of reading comprehension.[9]
4. Use simple, direct, unambiguous words. Make the stem and options as brief as possible.
5. Ask the candidate to choose the 1 best answer, rather than the correct answer; the cognitive processes required to make such a choice differ.[8]
6. Avoid topics that are controversial or debatable.
7. If there are differences in acceptable techniques or methods for doing something, avoid questions about specific techniques.
8. Difficult items are acceptable if they test at a desired cognitive level; make certain the difficulty is not due to poor wording.
9. Avoid ambiguous and misleading items.

Often a scenario can be included with several items to measure a candidate's knowledge pertaining to a situation. These are excellent items for determining ability to apply academic learning to real-life situations. Make certain, however, that 1 item does not answer or provide clues to another item. Keep each item independent of the others. Make the items universal by asking "What should be done?" rather than "What would you do?" What 1 person might do in a given situation might be entirely different from what another would do, and both actions might be correct or at least not harmful to the public. Items should be free of language that is offensive to a particular race, sex, ethnic group, or religion. Each item must be free of bias.

Suggestions for the stem

When constructing a stem, clearly define the problem to which the candidate must respond.[9] All information necessary for the candidate to identify the intent of the item should be presented in the stem. Candidates should be able to determine the type of response expected without having to read all of the options. Ideally, the candidate should be able to answer the question without looking at the options. The following is an example of a clearly defined stem:

Q. When determining the optimal time to proceed with surgery at the cessation of cranial growth, the most reliable method is

a. Hand wrist radiograph

b. Chronological age

c. Serial cephalometric radiograph superimpositions

d. Dental maturation

(answer: c)[10]

The stem should include as much of the item as possible so that words are not repeated in the options. Lengthy options can be distracting and confusing, especially when the information is repetitious. The following improperly constructed stem is followed by a properly structured alternative.

Q. According to Andreasen, autogenic transplants

a. are successful 50% of the time

b. are successful 65% of the time

c. are successful 75% of the time

d. are successful 95% of the time

e. are successful 100% of the time.

(answer: d)[11]

Q. According to Andreasen, autogenic transplants are successful in what percentage of the time?

a. 50%

b. 65%

c. 75%

d. 95%

e. 100%

(answer: d)

Use clear, simple language; avoid difficult and technical vocabulary unless it is essential to the intent of the item. Also, unless abbreviations are universally recognized, avoid them. The stem should give the candidate enough information to answer, without extraneous information.[12]

Avoid negative items. They often measure something of limited consequence, and candidates frequently find negatively worded stems confusing. Negatively worded items might be appropriate when candidates must know the contraindications or what to avoid. It is better to test what a candidate knows is correct rather than what he or

she knows is wrong. If, however, it is necessary to use negative words, they should be underlined, capitalized, or otherwise emphasized to clearly distinguish them. Following is an example of a negative word in the stem.

Q. Each of the following is a classic esthetic feature of apertognathia except one. Which one is the exception?

a. Stomion-soft tissue menton distance decreased
b. Lip incompetence
c. Posterior crossbite
d. Excessive height of the lower third of the face
e. Anterior dental open bite
(answer: a)[13]

The stem should present information or a problem that has general applicability rather than being specific to a limited situation. Avoid items that measure a concept or idea that would require different responses based upon where a candidate lives or was educated.

Avoid stems that ask candidates to decide on the correct definition of a term. The intent of credentialing examination items is to test ability to apply classroom knowledge to on-the-job situations. Create items that present problems so that the candidate must know the definition of a term to select the correct answer to a more complex problem.

Include all qualifications needed to choose the right answer. Some candidates might have had different, but not improper, training. Avoid subjectivity in the items.

Suggestions for the options

Select and form the options with care. Each distracter should be just as reasonable as the correct answer. If the distracters are totally alien to the stem, candidates will have no difficulty selecting the correct response. The more plausible and reasonable the options, the more the item will measure a candidate's competency (if the differences among the key and distracters are not minuscule).

Make the distracters attractive to the uninformed.[2] This does not mean that the item is tricky; rather, it means that the item has been well constructed. Candidates should not be able to pass an examination because the correct answers are obvious; they should pass because they can differentiate among the various options and select those that are correct. The following is an example of an item with effective distracters:

Q. At what developmental age should myofunctional therapy first be considered in a patient with a tongue thrust but without a speech problem?

a. Primary dentition
b. Early mixed dentition
c. Late mixed dentition
d. Permanent dentition-postpubertal
(answer: d)[14]

The following suggestions should assist in writing plausible distracters.

- Use common misconceptions.
- Make the distracters and the key similar in terminology and length.
- Avoid distracters that are opposites of the key.
- If mathematics is involved, use options that candidates would obtain if they worked the problem incorrectly.
- Vary the location of the correct answer.[12]
- Avoid using definitive words (always, never) in the distracters and generalities in the key.
- Use "good-sounding" words in both the distracters and the key.
- Make distracters similar to the key, but avoid fine distinctions that are of no practical significance.
- Arrange the options logically. Put numbers (years, weights, time, and so on) in ascending order and arrange multi-word or sentence options from shortest to longest. This method of sequencing assists candidates as they take the examination, but it also helps the writer become aware of placing the longest option (often the correct answer) at the end.
- Avoid using "all of the above" and "none of the above" as a correct answer. Both are overused and do not provide appropriate information about the candidate's competency. By using "all of the above," a candidate could answer the item correctly without really knowing the necessary information (eg, a candidate knows that distracters A and C are correct; therefore, "all of the above" must be correct.) The argument could also be made that regardless of which option the candidate selects, he or she would be correct because all the options are correct. Test-wise candidates will be certain to mark "all of the above" if they are uncertain about the correct answer because this option is typically correct. If "none of the above" is the correct answer, it is impossible to determine whether the candidates really know what is correct because they were not required to select a specific answer.
- Make all options grammatically correct and parallel in form with the stem. Avoid mixing stems that take a plural response with singular distracters; a test-wise candidate will be able to select the right answer if only 1 option is grammatically correct.
- Ensure that all options maintain a similar relationship to the concept in the stem. If the stem seeks a "how" response, all options should provide information on "how" and not on "when" or "where."
- Avoid trivial details or tricks, such as minor errors in a distracter (98.6°C instead of 98.6°F).

- Don't make the answer a direct quotation from a textbook. Write the correct answer and the distracters in your own words. Although all items must be referenced to readily accessible textbooks or journals, they should be written to measure a general or global concept rather than a specific sentence or paragraph in a book.
- Avoid responses that overlap each other or mean the same. For example, if the correct answer is "less than 2 years," then "less than 1 year" could also be correct. If options that mean the same are included, candidates can eliminate them because they know that only one option is correct. Options should be independent and mutually exclusive.
- Make certain that the key is truly correct. Candidates are told to select the best response, so there should be only 1 irrefutably correct response.

ITEM REVIEW

Once you have written an item, or if you are asked to review previously written items, ask yourself the following questions. In all instances, your response should be "yes." If you cannot respond positively, the item should be reconsidered.

1. Does the item measure a concept that is critical for the protection of the public or client?
2. Does the item measure only 1 concept?
3. Is the item directly related to the task and KSA for which it is intended?
4. Is the item presented in a straightforward manner?
5. Is the meaning of each item clear and precise?
6. Is the item current and correct?
7. Is the item written at the appropriate reading level?
8. Is the cognitive level identified correctly?
9. Have offensive or stereotypical references been eliminated?
10. Is the item free of debate, controversy, or dependence on a particular philosophy?
11. Do difficult items require a higher level of reasoning rather than requiring knowledge?

CONCLUSIONS

The ultimate purpose of an examination, including the ABO Phase II written examination, is to determine the candidate's competency for public protection. The collateral intent is that competency will lead to proficiency and eventually to excellence in orthodontic specialty care. Development and submission of questions for the Phase II examination should help to realize these goals. Questions that have insufficient relevance to optimal care delivery are inappropriate for the Phase II written examination. Our entire testing process must be directed toward maintaining Dr Ketcham's original objectives: "to elevate the standard of the practice of orthodontia" and "to protect the public against irresponsible and unqualified practitioners"[6] The public's trust remains in our hands.

REFERENCES

1. Schultheis NM. Writing cognitive educational objectives and multiple choice questions. Am J Health Syst Pharm 1998;55: 2397-2401.
2. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 3rd ed 2001.National Board of Medical Examiners; Philadelphia:
3. Stiggins RJ. The design and development of performance assessments. Educ Meas: Issues Prac 1988; 6:33-42.
4. Rogers WT. An empirical comparison of three-and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. Educ Psychol Meas 1999;59:234-47.
5. Barnette JJ. Effect of stem and Likert response option reversals on survey internal consistency: if you feel the need, there is a better alternative to using those negatively worded stems. Educ Psychol Meas 2000;60:361-70.
6. Cangialosi TJ, Riolo ML, Owens SE Jr, Dykhouse VJ, Moffitt AH, Grubb JE, et al. The American Board of Orthodontics and specialty certification: the first 50 years. Am J Orthod Dentofacial Orthop 2004;126:3-6
7. Test item development guide, National Board of Dental Examiners, Joint Commission on National Dental Examinations.
8. Bell TW. 14 rules for writing multiple choice questions. Brigham Young University 2001 Annual University Conference.
9. Guide for writing effective multiple choice items, Joint Commission on National Dental Examinations.
10. Proffit WR, White RP, Sarver DL. Contemporary treatment of dentofacial deformities. Saint Louis: Mosby; 2003. p. 533.
11. Northway W, Konigsborg S. Autogenic tooth transplantation. The "state of the art." Am J Orthod 1980;77:146-62.
12. Kehoe J. Writing multiple choice test items. Practical Assess Res Eval 1995;4(9). Available at: http://pareonline.net/getvn.asp?v=4&n=9). Accessed on May 22, 2005.
13. English J. Early treatment of skeletal open bite malocclusion. Am J Orthod Dentofacial Orthod 2003; 121:563-5.
14. Proffit WR, Mason RM. Myofunctional therapy for tongue-thrusting: background and recommendation. J Am Dent Assoc 1975;90:403-11.