Developing Test Items for Course Examinations

IDEA Paper #70 • June 2018

Thomas M. Haladyna • Arizona State University

Abstract

Writing multiple-choice test items to measure student learning in higher education is a challenge. Based on extensive scholarly research and experience, the author describes various item formats, offers guidelines for creating these items, and provides many examples of both good and bad test items. He also suggests some shortcuts for developing test items. All of this advice is based on extensive scholarly research and experience. Creating valid multiple-choice items is a difficult task, but it contributes greatly to the teaching and learning process for undergraduate, graduate, and professional-school courses.

Keywords: Multiple-choice items, selected response, test-item formats, examinations

In this article, I address how to develop test items to accurately measure student achievement in higher-education classes. There are two types of formats from which to develop test items: selected response (SR) and constructed response. Both are very useful. The SR formats include the popular multiple choice (MC) and the true-false (TF), among others.

Constructed-response formats include student performance, research paper, essay, experiment, oral report, observation, simulation, creative work, project, portfolio, and demonstration. I focus here on SR test items. For a comprehensive treatment of how to create highly effective constructed-response items, see Haladyna and Rodriguez (2013, pp. 189–260).

Why Do We Need to Develop Test Items?

A *summative test* is one that counts toward a student's course grade. Summative tests necessarily select their content from a large domain of knowledge, and a summative test is typically very comprehensive. A *formative quiz* is administered to students during the course so that they can chart their progress. Formative quizzes are intended mainly to help students learn but can also count toward a course grade.

What Kind of Content is Appropriate for SR Test Items?

Two important services to students in the teaching-learning process are (a) the clear identification of the course content that they need to learn and (b) the communication of that course content to them. If, as an instructor, you do nothing else, simply providing these two services is a very effective way to increase student learning. Accurately testing their grasp of course content is challenging but very satisfying for the instructor and the students.

Below are four categories of course content. Depending on what your teaching objectives are, these four categories of content figure importantly in your test items and testing. These four categories are hierarchical. Facts and concepts are fundamental in learning; principles and procedures represent more complex forms of content.

Fact

A fact is simply a true statement that few in your field would dispute:

1. A kilometer is 0.625 of one mile.
2. A period, exclamation point, or question mark ends a sentence.
3. Time is infinite.

Most educators would argue that we do too much testing of factual knowledge, and that learning facts alone does not constitute a very sound education.

Concept

A concept is an idea that is characterized by a definition, distinguishing characteristics, and examples and counterexamples. For instance, in the discipline of writing, concepts include expository, narrative, persuasive, and creative styles, as well as the mechanics of punctuation, spelling, and grammar, and the architecture of word, sentence, and paragraph. Teaching, learning, and testing concepts are important, because once students master concepts, they can apply what they have learned to more complex analysis.

Principle

"A principle is an accepted rule of action or conduct, a fundamental, primary, or general law or truth from which others are derived, a fundamental doctrine or tenet" (Haladyna & Rodriguez, 2013, p. 35). Some common principles include (a) "the theory of evolution is well supported by evidence," (b) "a paragraph should begin with a topic sentence," (c) "warm air rises," and (d) "exercise increases muscle mass." As with concepts, teaching and learning principles fosters more sophisticated thinking. Applying principles to complex problems and situations is encouraged at all levels of education. Testing students' understanding of principles is challenging, but it can be done, using the techniques that I will describe herein. Teaching, learning, and applying principles to complex problems and situations represent a higher form of thinking that is encouraged at all levels of education, particularly higher education.

Procedure

A procedure is an observable mental or physical action, such as hitting a golf ball, performing a calculation, microwaving popcorn, or washing clothes. We do not use SR to test the competent execution of a procedure. If we want to know if a student has learned what steps are involved in a procedure, however, SR testing is appropriate. For example, "What steps do you follow when washing a load of clothes?"

To summarize, these four categories of content help you identify what you teach, what students learn, and what you test. This is only the beginning of the test-development process. The next step involves assessing the mental complexity of what students learn and what you will test, also known as *cognitive demand*.

Cognitive Demand

For the most part, undergraduate- and graduate-level teaching is concerned with higher forms of thinking. Memorizing facts and simply understanding concepts, principles, and procedures is not enough. We want students to learn how to combine knowledge in complex ways to achieve some purpose. We might call the highest cognitive demand *problem solving*, *critical thinking*, or some similar term. Regardless of which term we use, the main idea is that knowledge is applied analytically to respond to a situation.

Recall

We expect students to recall or recognize, often verbatim, facts, concepts, principles, or procedures presented previously in class or in class readings. As stated earlier, recall is overemphasized at all levels of instruction and is the least important technique for educating students in any subject matter or profession. Examples of recall questions include the following:

- In what year was the Treaty of Ghent signed?
- What is Toulouse-Lautrec's first name?
- What is the Roman numeral for 1,000?

Comprehend/Understand

The meaning of a concept, principle, or procedure is an important part of teaching, learning, and testing. It is very easy to identify a verbatim definition or statement about a concept, principle, or procedure. Writing a test item to measure comprehension/understanding requires the item writer to produce novel examples or paraphrases of definitions of that concept. Here are some sample questions that test for comprehension:

- Which of the following is an example of onomatopoeia?
- What factors affect population density?
- What distinguishes an alluvial plain from other geologic features?
- What is characteristic of balance in furniture arrangement?

Application

The application of knowledge requires the student to use knowledge in a complex way, such as in a vignette, problem, or situation. For example:

- You purchase an item online with a 30% discount. Tax is 9%. The item is listed as $123 before the discount. How much will the item cost?
- Which of the following broker-dealing rules is in violation of Rule 15(c)3-1?
- You are planning a trip from Los Angeles to Denver in your electric-powered car. Which factors are most important in your plan?

From here on, I will focus on useful SR formats and guidelines for item development, and some shortcuts. You can save test items electronically in labeled folders for quick retrieval, to be used and reused on your quizzes and tests. A large "item bank" of this nature is a huge asset for any course that you teach.

Disclaimer

Cognitive demand is an elusive concept. A complex test item may prove to be a matter of simple recall for a highly experienced and intelligent student. A simple item might be cognitively complex for a student new to the subject matter. When you create an item, the intended cognitive demand is your best guess at what type of behavior a typical student will require to respond to the item. Generally, test items with a higher cognitive demand (comprehension/understanding and application) are preferable.

SR Formats

Besides the standard MC item, there are a variety of SR formats from which to choose. SR items are easy to write, and they very effectively measure comprehension and understanding and application of knowledge. In the following sections, recommended items (good examples) are presented normally, whereas items that are not recommended (bad examples) are shown ~~crossed out~~.

Multiple Choice

The MC item has a stem (question or prompt) and a set of choices (answers). There are two categories of MC: the question format and the sentence-completion format. The question format is more direct, but either format is acceptable. For example:

Which is the most effective treatment for infected hair follicles? [item stem]

A. topical medication [distractor—wrong answer]
B. astringent soap [distractor—wrong answer]
C. antibiotics [correct answer]

The most effective treatment for infected hair follicles is [item stem]

A. a topical medication. [distractor]
B. an astringent soap. [distractor]
C. an antibiotic. [correct answer]

Theory, research, and practicality indicate that three choices are sufficient (Haladyna & Downing, 1993; Rodriguez, 2005). Writing that fourth or fifth option is very difficult, as anyone who has written MC items will admit. If it is convenient or logical to have four options, then by all means add a fourth. But often, at least one of the additional distractors in a four- or five-option test item is so implausible that no student would choose it. Adding a useless distractor wastes your time as well as the students'. Three options are, therefore, strongly recommended.

Alternate Choice (AC)

In some circumstances, two options may suffice. Several researchers have discovered that with very high-achieving students (e.g., those you might find in a university or graduate or professional program), the most effective MC item consists of two options: the right answer and a plausible distractor (Ebel, 1981; Lord, 1977). For example:

When the water-to-cement ratio is high, what is the strength of the concrete?

A. high
B. low

Complex Multiple Choice (CMC)

This format is very popular. It is *not* recommended:

Which of the following rivets is used in steel framing?

1. button heads
2. knuckle heads
3. countersunk
   A. 1 & 2
   B. 1 & 3
   C. 2 & 3
   D. All of these

A comprehensive review of the CMC by Albanese, Kent, & Whitney (1977) led to these conclusions. First, these items tend to be very difficult, as reported by researchers. Second, the correct answer is often arguable. Partial knowledge of the correctness of the three foregoing choices (1, 2, and 3) may create conjecture about which combination is correct. Third, this format takes more time to develop, and yet each item still counts for only one point on a test. Fourth, a CMC item takes longer to read. Fifth, every item should accurately discriminate between high and low achievers, as measured by the *discrimination index*. CMC items prove to be less discriminating. Low discrimination negatively effects the reliability of test scores.

True-False (TF)

The TF format has a poor reputation, but undeservedly so. TF items are simply declarative statements that are absolutely true or false; there can be no doubt about the correct answer. The benefit of using this format is that true and false statements are easy to write. Thus, you can create and use many true-false items effectively. The tests are easy to administer, and, because TF items can be answered rapidly, a TF test can contain more items than other formats, making test scores more reliable. However, the floor of the scale for true-false is 50%, so that must be taken into consideration. Astudent scoring 50% on this kind of test, therefore, knows nothing about the content being tested. The following are examples of TF:

- A remote is a broadcast transmitted too faintly to be easily received far away.
- The most popular brew in the United States is lager.
- The first thing a mechanic does when a transmission does not work correctly is to conduct a pressure test.

Multiple True-False (MTF)

The MTF is a variation of the TF, with many attractive features. It begins with a vignette or introduction, followed by a series of between 4 and 30 statements, scored true or false. The benefit of this format is that writing, administering, and scoring responses is easy. Again, scores can be very reliable. To wit:

Climate scientists have identified factors contributing to the warming of the oceans. Which of the following factors have been identified by these scientists as contributing to ocean warming, and which factors are not related to ocean warming? Mark A if related; mark B if unrelated.

1. ____carbon dioxide produced from internal combustion engines
2. ____normal changes in a weather pattern occurring every several thousand years
3. ____methane from animal waste
4. ____intense sunlight
5. ____coal-fired plants
6. ____solar electric panels
7. ____rebound from greenhouse gases
8. ____CFCs
9. ____population growth and density
10. ____tree harvesting
11. ____the variation in distance from Earth to the sun
12. ____industrial activities

Testlet

One of the most effective ways to measure higher-level thinking is the testlet, a minitest that consists of introductory material followed by a set of 3 to 12 test items of any SR format. In most instances, I use the MC format. Generally, the testlet should occupy no more than two pages in a test booklet. The introductory material might be a problem, a vignette, a reading passage, graphical material, a performance, a chart, art, poetry, a video, a cartoon, an experiment, or examples of poor writing that need editing. Writing an effective testlet demands creativity, time, and effort. However, the rewards are rich. Testlets can and should be reused, but you can vary them systematically to generate families of testlets, increasing your supply for

future formative and summative testing. The use of the testlet has spread to nearly all testing programs and to all fields of study because of its ability for measuring higher level thinking. I will thus address the testlet in detail and describe several different types.

Reading-Comprehension Testlet

Because literature, literary analysis, and reading for comprehension are so important in most fields of study, the reading-comprehension testlet is very popular and has the most extensive history, with very effective techniques being honed over time. Instead of providing a complete example of a reading-comprehension testlet here, with introductory material, I will share a device known as *item shells*. These are item stems that have been used successfully in the past (Haladyna & Rodriguez, 2014, pp. 144–146); I will revisit item shells later in this article. The introductory material should present information that accommodates test items that measure comprehension/understanding and the application of knowledge, instead of simple recall, and should ideally run no longer than a single page.

The following is a set of MC item stems that could be used to prepare a set of items. Each stem has a specific focus. As recommended previously, three choices (a correct answer and two distractors) are sufficient. See Haladyna and Rodriguez (2013, pp. 77–78, 82–84) for more information about and examples of this technique.

Main Idea

- The main point of the author's work is …or What is the author's main point?
- The main idea of the passage is…or What is the main idea of this passage?
- The purpose of the passage is…or What is the purpose of this passage?

Slant of the Author

- What is the motivation of this article?
- How does the author slant the article?
- What is the point of view of this author?
- What does the author think about his main idea [or point]?
- Why was this article written?

Explicit References

- Which statement below agrees with the intent of the article?
- Which statement below agrees with the main idea of the article?
- Where does the story take place?

Inference

- What can we infer from this article?
- Which of the following would the author least likely [*or* most likely] recommend?
- How did the author feel about this issue?
- What would happen if [some action was taken *or* some consequence occurred]?

Themes/Arguments

- Which statement below best summarizes [a paragraph]?
- Which is the main point of [any paragraph]?
- Which statement expresses an opinion?
- What constitutes a persuasive argument in this article?
- What theme pervades this article?

Specific Language

You may wish to test students' understanding of the following literary devices, using examples: alliteration, allusion, caricature, cliché, foreshadowing, hyperbole, idiom, imagery, irony, metaphor, metafive, motif, onomatopoeia, oxymoron, paradox, personification, pun, rhetorical question, sarcasm, simile, or tone. These sample item stems show the richness and variety introduced by a reading-comprehension testlet:

- As shown in line 65, . . . ?
- In line 22, what does the word [ ] mean?
- In line 23, what does the phrase [ ] mean?
- Which is an example of an idiom?
- On line 24, to what does the pronoun [ ] refer?

Mathematics Problem-Solving Testlet

Although this example is not college- or university-level material, it *briefly* shows how to present mathematics problem solving in a testlet:

> Luke works at a convenience store. His current pay is $9.25 per hour. As of January 1, his pay will be increased to $10.50. He works 10 hours a week. He has to pay 10% in taxes and Social Security.

Only item stems are presented here:

1. How much does Luke earn each month in the previous year [*or* new year]?
2. How much would Luke earn each week in the previous year [*or* new year]?
3. About what percent is his pay increase from the old year to the new year?
4. If Luke is planning to buy a cell phone, and he plans to use the amount of the pay increase, how many months will it take him to save enough money to buy a $600 used cell phone?
5. How much in taxes and Social Security did Luke pay each [month *or* week] in the previous [*or* new] *year?*

Most of the preceding items test the ability to perform simple calculations and to work with percentages. Item 4 is the most complex, and it requires a series of correct calculations. This item set could be polished by adding more stems. Also, the characters and numbers can be varied systematically to expand the number of testlets available. This simple mathematics-problem item set has great versatility and potential for testing both simple and complex types of student learning.

Science Problem-Solving Testlet

The American College Testing Program (ACT) contains the best examples of science problem-solving testlets. Because of copyright considerations, these items cannot be presented here, but many good testlets are available on the ACT website: http://www.act.org/content/act/en/products-and-services/the-act/test-preparation/science-practice-test-questions.html?page=0&chapter=0

A typical science problem-solving testlet includes the following elements:

- introduction—a scientific study, or two competing passages with different views
- the results of a study (study 1)
- the results of another study (study 2);
- perhaps a table, chart, or graph of results for either study . . .

. . . all followed by a series of three to six MC test items; for example:

1. What does study 1 [or 2] suggest?
2. How do the results of study 1 differ from study 2?
3. What may have accounted for this difference?
4. How did the experimental designs differ?

Also, items can address specific issues or results of a study and call for a demonstration of understanding or comprehension. They can also test for the understanding of scientific concepts.

Figural/Graphical

In nearly all fields of postsecondary education, charts, tables, graphs, illustrations, artwork, musical works, photos, and other visual material can be presented for analysis. The following example presents a scenario and a table. The student is asked to analyze the table and respond to a variety of questions. The items are suggestive but can be expanded to include other types of analysis. Also, the table can be varied to include any state (e.g., Ohio), which increases the bank of testlets by a magnitude of 50. Of course, this quantity is probably unnecessary, but it shows the potential for increasing the number of available testlets and also for choosing which item stems best fit your instructional intents.

Arizona

AVERAGE WIND SPEED—MPH

STATION	ID	Years	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
CASA GRANDE AP AWOS	KCGZ	1996–2006	4.7	5.2	5.6	6.5	6.1	6.4	6.5	5.9	5.5	5.5	4.8	4.8	5.6
CHANDLER-WILLIAMS GTWY	KIWA	2001–2006	5.2	5.6	6.6	7.4	7.3	7.7	7.6	7.0	7.0	6.2	5.8	5.1	6.6
DOUGLAS AIRPORT ASOS	KDUG	1996–2006	7.2	8.1	9.0	10.3	9.3	9.4	8.2	6.9	7.2	7.6	7.0	6.8	8.0
FLAGSTAFF AIRPORT ASOS	KFLG	1996–2006	6.2	7.1	7.1	8.9	8.0	7.8	5.6	4.4	5.4	5.8	6.2	6.6	6.6
GILA BEND AIRPORT	KGBN	1996–2006	6.9	7.4	7.8	9.2	8.5	9.0	8.2	7.6	7.5	7.3	6.5	6.6	7.7
GLENDALE-LUKE AFB	KLUF	1996–2006	7.3	8.0	8.7	9.8	9.6	10.1	9.9	9.3	8.4	7.9	7.4	6.9	8.7
GRAND CANYON AP ASOS	KGCN	1996–2006	5.8	6.8	6.8	8.3	7.7	7.7	5.6	4.9	5.9	5.9	5.7	5.9	6.4
KINGMAN AIRPORT ASOS	KIGM	1996–2006	7.9	9.0	9.8	11.3	11.5	12.3	10.9	10.3	9.2	8.7	7.7	7.8	9.7
NOGALES AIRPORT ASOS	KOLS	1999–2006	5.1	6.2	6.8	7.7	7.1	7.4	5.7	4.6	5.4	5.7	5.0	5.1	5.9
PAGE AIRPORT ASOS	KPGA	1996–2006	3.3	4.2	5.4	7.0	6.7	6.6	5.9	5.4	5.1	4.4	3.5	3.0	5.0
PHOENIX–DEER VALLEY AP	KDVT	1996–2006	5.0	5.8	6.5	8.1	7.7	8.0	8.0	7.7	7.2	6.5	5.6	5.1	6.8
PHOENIX–SKY HARBOR AP	KPHX	1996–2006	4.9	5.6	6.4	7.6	7.3	7.6	7.6	7.2	6.4	5.7	5.1	4.6	6.3
PRESCOTT AIRPORT ASOS	KPRC	1996–2006	6.3	7.2	8.1	9.3	9.0	9.2	7.7	6.8	7.2	7.1	6.4	6.3	7.5
SAFFORD AIRPORT ASOS	KSAD	1997–2006	7.1	8.2	8.8	9.9	9.6	10.0	9.6	8.2	8.0	7.8	7.3	7.4	8.4
SCOTTSDALE AIRPORT ASOS	KSDL	1996–2006	3.0	3.9	4.5	5.7	5.4	5.8	6.0	5.5	5.0	4.1	3.4	3.0	4.6
SIERRA VISTA–FT HUACHUC	KFHU	1996–2006	10.2	11.5	11.9	12.9	12.1	11.8	9.8	8.6	8.9	9.4	9.1	9.2	10.5
ST JOHNS AP ASOS	KSJN	1999–2006	6.4	7.7	8.8	10.8	10.0	9.6	8.1	6.8	7.1	6.6	6.5	5.6	7.8
TUCSON AIRPORT ASOS	KTUS	1996–2006	6.7	7.2	7.7	8.3	8.1	8.3	8.0	7.5	7.6	7.6	7.0	6.8	7.6
TUCSON–DAVIS MONTHAN AF	KDMA	1996–2006	7.5	7.9	8.1	8.7	8.5	8.9	8.9	8.3	7.9	7.9	7.6	7.6	8.1
WINDOW ROCK AP ASOS	KRQE	1998–2006	4.4	5.5	6.4	8.4	7.7	7.2	5.3	4.4	4.9	4.7	4.6	4.1	5.6
WINSLOW AIRPORT ASOS	KINW	1996–2006	6.6	7.7	9.0	11.0	10.6	10.4	8.9	7.9	7.8	7.3	6.7	6.4	8.3
YUMA MCAS–INTL AP ASOS	KNYL	1996–2006	7.1	7.2	7.6	8.7	8.2	8.6	9.1	8.6	7.2	6.7	6.5	7.4	7.7

Wind can be used to generate electricity. Your company is studying potential sites in Arizona.

- Given that Arizona has the greatest need for electricity in the summer months (June, July, August) . . .
- Which city has the least potential for generating electricity using wind?
- Which location has the greatest [*or* least] potential for electricity generation using wind in the summer [*or* entire year]?
- Your company has decided that 8.0 is the lowest wind speed for any installation; which of the following cities qualifies [*or* does not qualify]?
- Given what climate scientists have predicted, what is the most likely outcome for electricity generation using wind in Arizona [*or* some other location]?
- Which is true regarding the Kingman and Sierra Vista sites for wind generation of electricity? (This could be presented in an MTF format.)
- You have been assigned to present a plan for future installation. What is your first step?
- What is the most important element in your plan?

These item stems represent a small sample of possible MC test items that might be generated for a testlet. As you can use information from any state, the number of test items generated is more than you will ever need. These items can be used for practice, formative quizzes, or summative tests.

Guidelines for Creating SR Items

The opinions and wisdom of textbook authors on testing, backed up by much research, led to the development of a set of item-writing guidelines (Haladyna & Downing, 1989). These guidelines have been revised several times. The latest version is extracted from Haladyna & Rodriguez (2013, p. 91). For many of these guidelines, examples of a guideline violation are presented. Such bad examples are crossed out.

Content

1. Each item should be based on one type of content and one cognitive demand.

Sometimes a test item will require two steps to reach the correct answer. If the student misses the item, you will not know which step was the problem; therefore, every item should have a singular focus.

1.1. ~~What is a synonym for the word charity?~~
1.2. What is a synonym?
1.3. Below are pairs of words. Mark A if a pair is an example of antonyms or B if it is an example of synonyms.

The first item requires two mental steps. First the student has to know what a synonym is. Then the student has to identify the synonym for the word. The second item merely establishes that the student knows the meaning of synonym. The third item focuses on distinguishing between antonym and synonym, after it has been established that students have had the opportunity to learn the meaning of both terms.

2. Use novel material in your test item.

Do not use language verbatim from atextbook or class discussion; paraphrase if you are testing a student's comprehension or understanding. The following example uses the MTF format. Note that the choices (not shown) are meant to be original and not encountered in a textbook, lecture, or other source that students have previously seen.

2.1. Which sentences below are examples of sarcasm in literature?

3. The content of each item should be independent of other content.

One item should not cue another item. Note that the answer to the first of the two following items provides a clue regarding the second item. If hydrocollator is adjacent to the second item, students may be cued that C is the correct choice for item 1. This cueing can be eliminated by using contrast bath and Russian bath in other items on the quiz or test.

3.1. ~~Which form of hydrotherapy is most effective?~~
~~A. contrast bath~~
~~B. Russian bath~~
~~C. hydrocollator~~
3.2. ~~With the use of a hydrocollator, what is the most severe limitation?~~
~~A.ineffective with obese patients~~
~~B.transfers heat less effectively than air~~
~~C. less effective with patients who have acute pain~~

4. Avoid too-specific or too-general content; emphasize important content instead of trivial content.

The following items, 4.1 and 4.2, are too general and too specific respectively. Item 4.3 deals with trivial content—memorizing a date. Item 4.4 is so general and expansive that it would be difficult to develop a right answer and plausible distractors.

4.1. ~~Which strategy is most effective in achieving world peace?~~
4.2. ~~In what year was the Treaty of Ghent signed?~~
4.3. ~~What is Rimsky-Korsakov's first name?~~
4.4. ~~What is the meaning of life?~~

5. ***Avoid opinions, unless so qualified.***

Facts and opinions are always contrasted; students learn early in life that there is a difference. Testing should generally stick to facts, though occasionally knowledge of opinions is tested. There is a right and a wrong way to do this.

5.1. ~~What element gives balance to a room?~~
A. a window on every wall
B. nearly square dimensions
C. furniture of different heights

5.2. According to the National Institute of Interior Design, what feature gives balance to a room?

A. a window on every wall
B. nearly square dimensions
C. furniture of different heights

Example 5.1 shows a brief first stem, calling for an opinion that can be argued by students, and confusion may arise, given their different points of view. The second stem, 5.2. qualifies the opinion's source, using the same three options.

6. ***Avoid trick items.***

Some items intentionally deceive the student into choosing the wrong answer. Trick items should always be avoided. First, deception in teaching and testing accomplishes nothing positive. Trick items create anxiety in students, and test anxiety is a very common problem for all test takers that we wish to reduce. Trick items do not contribute toward validly measuring what a student has learned. The following examples are provided more for amusement, but they illustrate the point.

6.1. ~~A plane crashes on a river between two states. Where are the survivors buried?~~ [Nowhere; they are survivors.]
6.2. ~~How do the British celebrate the Fourth of July?~~ [Just like any other day.]
6.3. ~~Where do Panama hats originate?~~ [Chile.]
6.4. ~~July has 31 days. How many months have 28 days~~? [All months.]

Format

7. ***Items should be formatted vertically, not horizontally.***

Although horizontal formatting saves space on the page so that more items might fit, reading choices horizontally may be difficult or confusing. This type of formatting can induce anxiety in students, and, noted previously, test anxiety is a major problem among test takers of all ages. Vertical formatting is, therefore, recommended. All testing programs that I know of now use vertical formatting.

7.1. ~~Battery electrolyte consists of water and (A) lead sulfate, (B) sulfuric acid, (C) lead peroxide.~~

What is combined with water to form battery electrolyte?

A. lead sulfate
B. sulfuric acid
C. lead peroxide

8. ***All items should be carefully edited and proofread.***

Faulty grammar and typographical errors can be distracting and reflect badly on your instruction. (Examples will not be demonstrated here.) Test companies spend considerable resources to ensure that their items are perfect. Although you probably do not have the resources to spend on professional quality assurance, do what you can regarding editing and proofing, to benefit both yourself and your students.

9. ***The linguistic complexity of items should be appropriate for the students being tested.***

Increasing numbers of students in higher education come from countries where English is not the first language or may not be widely spoken. Those English-as-a-second-language learners face additional challenges when navigating assigned reading, lectures, and testing. In professional schools, the problem is exacerbated because graduates have to pass high-stakes certification or licensing tests that present the same challenge: How can you make tests fair to all students?

First and foremost, use language that is appropriate for the subject matter and the students being tested. Do not construct lengthy sentences and complex paragraphs as stems for items. Use the necessary vocabulary for the subject but try to keep it simple enough for students to grasp the intent of the test item. Jargon should be avoided. Acronyms can be used if they are universally recognized. Choose active voice over passive voice whenever possible. Enliven items by using individuals instead of vague references to "people." The crossed-out version of the following item precedes a version showing characters discussing a real-life problem.

9.1. ~~Blue smoke is coming from the exhaust pipe. A says a blocked cylinder head is the cause. B says the thermostat is stuck open. C says it is not a problem~~.

9.2. Seeing blue smoke coming from an exhaust pipe, Larry thinks it is a blocked cylinder head, Moe says it's a stuck thermostat, and Curly says it is not a problem. Who's right?

A. Larry
B. Curly
C. Moe

Writing the Stem

This section contains advice on how to write the item stem. Later in this article, considerable assistance is provided to improve stem development.

10. Minimize the number of words in the stem of the item; avoid window dressing.

You can evaluate the wordiness of any test item by simply counting the words. Students have to read these words and extract meaning within a limited time. A wordy test item takes more time to administer, and a less-wordy item counts the same as a wordy one—one point. Conciseness is a virtue, as long as you communicate the essence of the content you want to test for. The options can be brief yet still informative. The first of the following two examples expresses the intent of the item clearly in five words. The options can be brief yet informative.

10.1. Which person below displays maladjustment?

Here is a 136-word item based on material presented by Bragdon (1989, p. 160):

10.2. ~~A groundwater basin occupies an area of 11,000 acres. Throughout an ideal hydrologic cycle, the following vegetation has existed in the basin: alfalfa—3,000 acres; corn—2,000 acres; brush—6,000 acres. Rainfall in the basin has averaged 10 inches per year; outflow from streams, 4,000 acre-feet per year; and inflow from streams, 12,000 acre-feet per year. Subsurface inflow has averaged 800 acre-feet per year, and subsurface outflow has been insignificant. A water company has exported 150,000,000 cubic feet of water annually to lands outside the basin watershed. Assume the following annual consumptive use values: alfalfa, 4 feet; corn, 2 feet; brush, 0.6 feet. Specific yield of alluvial material is 13%. What is the average annual change in storage within the basin in acre-feet?~~

A. much less than 600
B. about 600
C. much more than 600

There is far too much material in this stem. We call this *window dressing*; much of it is irrelevant. However, this item need not be discarded entirely. The vignette in this item stem could provide a basis for a comprehensive testlet.

11. State the central idea in the stem, not in the options.

Sometime an item will have the opposite problem as that stated in guideline 10: the stem is brief, and the options are very long. The following item's options are too wordy, and there are too many of them. It concerns advice on how to grade students.

11.1. Which is best for assigning a student grade?

~~A. Whenever students are graded on a curve, a percentage will fail the course.~~
~~B. Raising grades from low to high is a good way to increase standards in a school.~~
~~C. As a teacher, you are better positioned to defend a grading system supported by your administration than on the basis of your perceived value to students.~~
~~D. The most meaningful form of reporting is percentage of points earned divided by the percentage of points possible.~~
~~E. Grades should be based on the results of standardized achievement tests as opposed to an instructor's test.~~

Another problem related to wordy options is the *unfocused stem*. We err on the side of economy when we reduce an item stem to a single word. The stem should represent the main intent of the item. The unfocused stem leaves the student wondering about the item's intent. The following example exemplifies this confusion.

11.2. ~~Edgar Allan Poe~~

~~A. was a writer, editor, and critic.~~
~~B. was mentally ill.~~
~~C. died from complications of pneumonia.~~

*12. **Use positive language; avoid negative terms such as "not", "except", or "false."***

The authors of textbooks on testing, as well as some research, suggest that the use of negative terms in the stem has a negative effect on student answers. The following item would be better presented without the not*, and with two noncharacteristics and one characteristic. However, if you must use* not *or some other negation, make sure that it appears in capital letters, and even boldface. That way, a student is unlikely to be confused.*

12.1. Which is NOT characteristic of the climate of Cabo San Lucas?

A. warm winters
B. frequent summer thunderstorms
C. fall typhoons

Another option is that items such as this might be converted to the MTF format, where the stem reads, "Which are characteristic of the climate of Cabo San Lucas?" Then follow with a series of three or more statements, some of which are characteristic and some of which are not.

Writing the Options

13. Distractors should be plausible wrong answers, not ridiculous or otherwise implausible choices.

Perhaps the most difficult part of developing these test items is thinking up the correct answer and several distractors. Distractors are supposed to lure students who have not learned the content that you have taught and are testing for. For a long time, item developers have continued to supply four or five MC options. Anyone with experience in item writing will attest that thinking up the fourth or fifth option is very difficult. For many reasons, Haladyna and Rodriguez (2013) have recommended three options. If you have to create only twodistractors, you have a better chance of writing a highly effective item. Plausibility is a difficult concept to explain or demonstrate. It may help to think of a distractor as representing a common student error.

14. With any of the SR formats, there should be only one right answer.

Purely by accident, you might write an item that contains both your keyed response and another option that is arguably correct. Students are quick to point this out. Most universities and colleges have testing services that conduct item analysis as a service to faculty. This service is strongly recommended. If item analysis is done, you will find that the frequency with which a particular distractor is chosen is revealing. If the majority of students chose a distractor instead of the correct answer, there is a problem. The item should be revised to eliminate the troublesome distractor in the future, and the best damage control in the present is to give students credit for the secondary correct answer. In professional testing programs, more than one right answer per item is very rare. In tests constructed by instructors, this type of error is more common.

*15. **The location of the right answer should vary randomly.***

Avoid patterns. Students are always looking for clues. The location of the correct answer should be balanced. For a three-option, 30-MC-item quiz, 10 items should be keyed A, 10 items should be keyed B, and 10 items should be keyed C. Professional testing programs try to achieve this outcome. For your quizzes and tests, balancing the location of the correct answer is not as important, but it is desirable.

16. If numbers are used, options should occur in numerical order.

Make sure place value is correct, as demonstrated in the following example, item 16.1. In the crossed-out set of options, the 6 in 60 and 9 in 90 are aligned with the hundreds column. As noted previously, students with some test anxiety might be confused about place value. Moreover, it is simply sloppy. Numbers should be properly aligned for the sake of clarity and the aesthetics of the test. Numbers should also always be presented in ascending order.

16.1. In a regular hexagon, what is the measure of any internal angle?

~~A. 60~~	A.	60
~~B. 90~~	B.	90
~~C. 115~~	C.	115
~~D. 120~~	D.	120
~~E. 150~~	E.	150

17. Options should be independent of one another; avoid options that overlap in quality or content.

Two reasons for maintaining independent options are (a) two choices might be correct if the overlapping part

is a correct answer, and (b) a clue might result that enables a test-wise student to guess correctly. Item 17a is an obvious example of a violation of this guideline. In item 17b, note that Upton Sinclair and Lincoln Steffens appear twice in the choices, but the other two people appear only once. Any test-wise student might select B.

17a. According to NOAA, what is the increase in world temperature for 2017?

~~A. between 1 and 3 degrees~~
~~B. between 2 and 4 degrees~~
~~C. between 3 and 5 degrees~~

17b. Which of the following figures in history should be considered muckrakers?

~~A. Teddy Roosevelt and Upton Sinclair~~
~~B. Upton Sinclair and Lincoln Steffens~~
~~C. Lincoln Steffens and Joseph McCarthy~~

18. Avoid terms such as "all of the above" and "none of the above."

There is a lively debate about the desirability of using these terms in an MC item. *None of the above* seems convenient for those test creators who are unable to think of a third or fourth option. If used, it too should be the right answer. If you must violate this guideline, use none of the above. Never use all of the above. Occasionally, you might see the option *I don't know*—a useless one, because the naive student will confess that he or she doesn't know instead of guessing. Most test-wise students know that if they can eliminate one or two plausible distractors, their chance of guessing the right answer improves. *I don't know* should never be used.

19. The options should be worded positively, not negatively.

This same advice was given for the stem (guideline 12); the use of negation is often confusing to students. It is best to maintain a practice of avoiding it both in the stem and the options. If negation is truly desired, consider the MTF format, where a serial list is presented and the student can choose between positive and negative examples.

20. Avoid clues that hint at the right answer.

<u>The length of each option should be about equal</u>. Sometimes the longest option is the right answer. This kind of error is easily detected and is found more often than you might think possible. A careless item writer will write the correct answer and hurriedly create two short distractors.

20a. Which of the following is the most effective treatment of plebnex?

A. rest
B. ~~anti-inflammatory medication, adequate hydration, and exercise~~
C. aspirin

<u>Avoid specific terms: *never, always, completely, absolutely*</u>. Such terms are so extreme that options containing them are seldom correct. Also, for the test-wise student, the use of these specific terms is often a clue that these are incorrect choices. If an option containing one of these words is in fact the correct answer, it may still trick some students into avoiding the option. For example:

20b. You are buying a car for a loved one. Which advice is valid?

~~A. Always get a large car with lots of air bags.~~
~~B. Never get a small car.~~
C. Look for a car with a good safety and reliability record.

<u>Avoid clang associations</u>. Terms in the following options echo terms in the stem. Item 20c has the clue *voltage* in the stem. If B is not the correct answer, then the item doubles as a trick to deceive the student.

20c. Which is in the low-voltage circuit of the ignition system?

A. condenser
B. ~~voltage regulator~~
C. spark plugs

<u>Avoid pairs or trios of options that might give students a *clue to* the right answer</u>. Item 20d has two choices that use the word shingle, leading the student to choose gravel.

20d. Which type of roof requires a felt underlay?

~~A. asphalt strip shingle~~
~~B. wood shingle~~
C. gravel

Avoid ridiculous or blatantly implausible distractors. For the sake of humor or out of carelessness, an item writer may write a good stem and a good correct choice but then include a distractor that would never be chosen. Sufficient research supports the idea that at least one fourth of all distractors are so implausible that fewer than 5% of students would choose them (Rodriguez, 2005). That 5% or less represents students who are likely guessing randomly. See the following:

20e. Among the leading astrophysicists in the world, you would include

A. Neil deGrasse Tyson.
~~B. Princess Moonbeam.~~
~~C. Elon Musk.~~

The grammatical structure and content should be parallel in the choices. As with other types of errors, a careless item writer may write a good stem and correct answer but then add distractors that do not connect grammatically with the stem:

20f. The definition of swatting refers

~~A.to making false 911 calls claiming someone has a gun.~~
~~B.is a criminal offense.~~
~~C.civil claim involving a false 911 call.~~

21. Avoid humor.

Although it can be fun to create witty items, some students can be distracted by them. But if you disagree with this principle, use humor, but only if it reflects your personality demonstrated in your class.

As you may have noticed, item writers can violate many of these guidelines and still create highly effective items. Although all advice proffered in guidelines is based on the collective wisdom of those who study item development, in some circumstances you may wish to depart from the guidelines for the sake of expediency or some other reason.

Are There Shortcuts for Developing Items? Yes! For quite some time, my colleagues and I have been experimenting with ways to speed up item creation. *Automated Item Generation* (Gierl & Haladyna, 2013) is dedicated to the science of quickly generating large numbers of items that measure understanding or the application of knowledge. What follows here is a distillation of several practical methods and emerging technologies; I briefly mentioned some of these shortcuts earlier.

The Item Shell

As noted in the section on item formats, the reading-comprehension testlet contains a good set of item shells. An *item shell* is a hollow item stem. It provides a prompt for writing your item. Many item developers experience writer's block, and item shells help them overcome that obstacle. The following generic shells are provided to give you an idea of what an item shell is. Note that they are MC item stems, usually in the form of a question.

1. Which is the best definition of [a concept, principle, or procedure]?
2. Which is an example of [a concept, principle, or procedure]?
3. Which principle applies to this situation?
4. Which procedure is appropriate for this situation?
5. What causes this effect?
6. What is the relationship between x and y?
7. What would happen if . . . ?
8. Which is the most [*or* least] important . . . ?
9. What is the difference [*or* similarity] between x and y?
10. What is the best way to . . . ?
11. Which term below best represents [or summarizes] ...?
12. What is the best way . . . ?

This is not an exhaustive list. For a more comprehensive treatment of this topic, see Haladyna and Rodriguez (2013, pp. 144–146). Although this format is useful for quickly generating items, one caveat is that the items in your item bank will all begin to look alike. Therefore, item shells can be used to get you started, but do not over-rely on this format. Creativity when phrasing items is still desirable, and variety is a good thing.

Testlet Generation

I like MC testlets because they are an effective way to model and test for higher-level thinking, and they can be designed to be replicated in virtually endless variety. I have mentioned how easy it is to change a scenario or vignette to generate new testlets while using the same set of items. Having taught introductory, advanced, and multivariate statistics to doctoral students, I have developed vignettes with quantities that can be varied. In other words, the MC items remain the same but the

vignettes are systematically varied to represent different research scenarios. Simple manipulations of these quantities change the item sufficiently to produce new vignettes where the same items are used. Such items can be used for study, formative quizzes, or tests, and instruction is truly integrated with how I quiz and test.

The following example comes from teaching a graduate-level beginning statistics course. There is a generic scenario, or vignette. Note that the values *x* and *y* can be altered, which changes the testlet. Also, the variables used in the correlation can be renamed. Thus this method produces many parallel testlets, some more challenging than others.

> Professor Thumbsucker has hypothesized that those students preferring boiled lima beans have a sour disposition. His study involved [x] students, using highly valid and appropriate measures of disposition and the degree of love for boiled lima beans. He collects data for analysis. He computes a product-moment correlation. The correlation coefficient is [y]. He is also making a causal inference that if he can improve lima-bean lovers' diets, their disposition might improve.
>
> Who wants sour-puss students in class?

Here are some sample item stems that could be used as part of the testlet:

- What is the independent variable? What is the dependent variable? Which variable is designated x? Which variable is designated y?
- What is a parameter?
- What is a statistic?
- What is the null hypothesis? What is the alternate hypothesis?
- Which alpha level should you use to test a hypothesis?
- Which hypothesis should you choose, directional or nondirectional? Why?
- Given the result and an alpha of [.10, .05, .01, *or* .001], would you accept or reject the null hypothesis?
- What is the practical significance of your result?
- How do you calculate the degree of practical significance?
- Should you create and study a scatter plot? Why?
- What is the power of your statistical test?
- If the reliability for the first measure is [some value] and the reliability of the second value is [some value], how does that affect the estimation of correlation?
- If your reliability was perfect, what size of correlation might you expect?
- What about making a causal inference? Which statement is most defensible?
- If you were going to redo the study, what would you recommend as most important?

Recycling Items

Textbooks may have supplemental materials that include test items. Generally, these items are very poorly written. Rather than use them as is, you can recycle them by retaining the content and following the guidelines presented in this article to create new stems and options. At the very least, they may generate ideas for additional items of similar content and cognitive demand. You can also find items online but take care to avoid copyright violations. You may use these items if you adapt them by transforming them in some way (modifying quantities, content, or format); that is not considered stealing.

What Happens to Items After They've Been Developed?

Validating items for professional testing programs is very expensive. These items pass through many steps in item development, including actual use by students followed by statistical item analysis. For the college, graduate, or professional course, such a process is not feasible. The best that you can do is to beg, borrow, adopt, adapt, and create items and then store them on your computer in an item bank. Any word-processing program can store items for quick retrieval to construct practicetests, formative quizzes, and summative tests.

Summary

Creating items for course quizzes and tests is very difficult but necessary. It is part of the teaching and learning process that provides students with feedback on their progress and serves as a basis for grading. The concepts, principles, and procedures in this article were mainly drawn from a long period of study and experience in item development for testing programs, and in helping teachers-in-preparation create highly effective tests. As we have seen, there are many useful tools to help you in this journey.

Tom Haladyna is Professor Emeritus at Arizona State University, where he served in teacher education. His experience includes work as a public school teacher, faculty member at two universities, test director at American College Testing (ACT), research professor in the Oregon State System of Higher Education, visiting scholar at the US Navy Personnel and Research Development Center, and National Assessment of Educational Progress visiting scholar at the Educational Testing Service (ETS). As an educational psychologist, Tom has been involved mainly in test and item development and validation. He has authored many books, journal articles, and papers. He has also led various faculty-development activities during his career. The basis for this article comes from his new book with Michael Rodriguez, Developing and Validating Test Items, which is in its fourth edition.

References

Albanese, M. A., Kent, T. A., & Whitney, D. R. (1977). A comparison of the difficulty, reliability, and validity of complex multiple-choice, multiple responses, and multiple true-false items. *Annual Conference on Research in Medical Education, 16*, 105–110.

Bragdon, A. D. (1989). *The book of tests*. New York: Harper and Row.

Gierl, M., & Haladyna, T. M. (Eds.). (2013). *Automated item generation*. NY: Routledge.

Haladyna, T. M., & Downing, S. M. (1989). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*, 51–78.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item. *Educational and Psychological Measurement, 53*(4), 999–1010. doi.org/10.1177/0013164493053004013

Haladyna, T. M., & Rodriguez, M. R. (2013). *Developing and validating test items*. NY: Routledge.

Lord, F. M. (1977). Optimal number of choices per item—A comparison of four approaches. *Journal of Educational Measurement, 14*, 33–38. doi.org/10.1111/j.1745-3984.1977.tb00026.x

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice test items: A meta- analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2) 3–13). doi.org/10.1111/j.1745-3992.2005.00006.x

T: 800.255.2757
T: 785.320.2400

301 South Fourth St., Suite 200
Manhattan, KS 66502-6209
E: info@IDEAedu.org
IDEAedu.org