Constructing Test Items:

Multiple-Choice, Constructed-Response, Performance, and Other Formats

Second Edition

Steven J. Osterlind
University of Missouri-Columbia

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

Chapter 5

Style, Editorial, and Publication Guidelines for Items in the Multiple-Choice Format

INTRODUCTION

When authors refer to style, they usually mean the expression of ideas in a smooth, orderly, pleasing manner. Each author develops an individual style of expression that allows for a personal presentation of his or her own thoughts and emotions. For editors, however, style connotes something different. Editorial style refers to the consistent use of a set of rules and guidelines. These rules and guidelines prescribe a consistent use of punctuation, abbreviations, and citations, a uniform and attractive format for tables, graphs, and charts, and a correct form for the many other elements that constitute written communication. Test-item writers are both authors and editors. As authors, they can express their ideas in novel, creative, and personal ways; as editors, they must be aware of and follow consistent editorial principles.

Chapter 4 focused on the role of item writer as author, describing the importance of good writing to producing meritorious test items and offering suggestions to improve one's writing style. This chapter addresses the editorial aspect of writing good items, introducing editorial rules and guidelines for the presentation of test items. These rules and guidelines specify punctuation, spacing, alignment, and the appropriate use of charts, tables, and figures in test items. Additionally, several specify how to correctly cite terms commonly used in items, such as titles, abbreviations, quotations, mathematical symbols, legal references, and the like.

Another, and important, section of this chapter describes the mechanics of how to prepare a test item for final publication. Publication does not only mean items that will be typeset and printed by a professional printing service; rather, it means the final presentation of a well-constructed test item, regardless of whether the product is professionally typeset, or as is more commonly done today, produced on a personal computer with a word-processing and page-layout or desktop-publishing program. For professional results, item writers must know when to use boldface type or italics, as well as what style and size of type are pleasing to the eye. These publication characteristics and many more such concerns are addressed in this chapter.

All of the remarks and examples in this chapter describe test items in the multiple-choice format. The following chapter presents parallel rules and guidelines as they apply to other item formats, specifically performance-based or constructed-response formats, true-false, short-answer or sentence-completion, matching, and cloze-procedure. Of course, most of what is described in this chapter will apply to all item types. However, since the focus of this chapter is on multiple-choice test items, it will also address issues uniquely related to items in that format, such as their advantages and disadvantages.

The following topics are covered in this chapter:

- understanding the multiple-choice item format
- advantages and disadvantages of the multiple-choice format
- editorial format for multiple-choice items
- correct use and placement of directions
- specialized rules for formatting multiple-choice items
- type characteristics and appealing page layout.

UNDERSTANDING THE MULTIPLE-CHOICE ITEM FORMAT

The multiple-choice format for test items is characterized by an item with a stem sentence or phrase that presents a problem or asks a question, followed by usually two to five response alternatives, one of which appropriately answers the stem. Further, as has been described in earlier sections of this book, the stem should generally be in the interrogative. The response

alternatives may be complete sentences, sentence fragments, or even single words.

Although many of the discussions throughout this book use multiple-choice items as examples, one realizes upon inspection of this format that there is not just one style for all multiple-choice items. In fact, the multiple-choice item can assume a variety of types, including absolutely correct, best-answer, and those with complex alternatives. Features of multiple-choice items that are absolutely correct, best-answer, and those with complex alternatives were discussed in the preceding chapter. Wesman (1971) identified eight types for the multiple-choice item, although all of them could easily be placed into one or more of the three categories mentioned.

ADVANTAGES AND CRITICISMS OF ITEMS IN THE MULTIPLE-CHOICE FORMAT

Strengths of the Format

The multiple-choice format for test items has a number of strengths that make it particularly well-suited for assessing mental attributes. Perhaps principal among its advantages is the fact that it offers more flexibility for assessing a diversity of content and psychological processes than can be garnered from other item formats. Nitko (1983) remarked: "Among the various types of response choice items, the multiple-choice item can be used to test a greater variety of instructional objectives" (p. 193). And, Haladyna and Downing (1989a) state that testing organizations and technical measurement experts prefer the multiple-choice format because its "sampling of content is generally superior when compared to other formats." This flexibility for accommodating a diversity of content means that the multiple-choice format can be used in a variety of assessment instruments, whether designed for educational or psychological assessment.

An aspect of the flexibility of the multiple-choice format is that, when well-written, the format allows assessment of complex and sophisticated mental skills, ones that would be high up on a list of cognitive abilities. Stated less precisely and more bluntly: good multiple-choice items can assess higher order thinking skills.

A further advantage of the multiple-choice format is that it allows for a precise interpretation which can lead to important evidence for content-related test validity. According to Haladyna and Downing (1989b), "the use of MC formats generally leads to more content valid test score interpretations."

The advantage of multiple-choice items for content-related evidence for validity may be clearly demonstrated by comparing the format with another response mode, the essay, Consider the fact that the multiple-choice item does not require examinees to write their responses, which can become long and elaborate. When responses are written, as in essay examinations, an examinee can mask his or her limited knowledge of the subject being assessed by producing writing that obfuscates the issue. On the other hand, when measurement is accomplished through a set of multiple-choice items, examinees are less likely to be able to bluff their way through content-related material. Of course, this statement presumes measurement error has been minimized (e.g., no guessing, etc.).

Considering just the positive features of the multiple-choice item format, then, a set of well-constructed items in this format can provide very reliable assessment of an examinee's true ability. The format is flexible, sophisticated, and simple to use.

Criticisms of the Multiple-Choice Format

Items in the multiple-choice format also have negative aspects, some of which are the obverse of positive aspects of the format. For example, one positive feature of the format, that examinees need not produce a long, written response, can also be construed as a negative characteristic. By prescribing that examinees merely select one choice from among a very few alternatives, the multiple-choice format inhibits examinees from expressing creativity or demonstrating original and imaginative thinking. The critics charge that thorough discussion of events and their meaning is discouraged by reducing knowledge to only what can be accommodated by simple statements, since that is what fits most conveniently into the format.

Some detractors of standardized testing take the criticism that multiple-choice items inhibit creativity and reduce all important knowledge to superficial facts a step further by claiming that knowledge itself is "vulgarized" by the multiple-choice format. This vulgarization of knowledge is

reified for these critics in the sense that multiple-choice items suggest there is one, and only one, correct answer to every problem.

However, this denunciation of the multiple-choice item is not universally accepted. In fact, there is overwhelming proof that valid interpretations of scores from tests containing multiple-choice items can be made. To review this proof, one need only examine the evidence for valid score interpretations from literally hundreds of well-constructed tests. Further, the interpretations of test's scores are not limited to simple interpretations. Many times the score's interpretations are complex and sophisticated, lending further evidence that the criticism is unfounded. In separate essays, Wood (1977) and Ebel (1972) offer thorough and cogent defenses against this criticism.

Of course, the critics' charge can become all too accurate if items are poorly written. Regretfully, too many poor items exist. One hopes, however, poorly-prepared items will become increasingly rare. In a very real sense, this entire book is aimed at reducing the chances for the trivialization of knowledge which can occur with poorly-constructed test items. Preparing good test items requires not only all of the knowledge and background discussed in the preceding chapters, but also a familiarity with editorial format for items, information presented in the following section.

EDITORIAL FORMAT FOR ITEMS

Description of Editorial Style

Editorial style refers to the consistent use of rules and guidelines for punctuation and abbreviations, a uniform and attractive format for tables and other graphics, as well as proper citation and appropriate use of titles, names, dates, symbols, and many other special characters. For example, the mountain range bordering California and Nevada is correctly referred to as the Sierra Nevada (not Sierra Nevada Mountains or the Sierras). The Stone Age is capitalized but the nuclear age is not. Legal cases are cited as (including italics) *Plessy v. Ferguson.* And, time is properly noted as 3:15 p.m. and 12:00 noon or 12:00 midnight.

When constructing test items, writers should routinely and uniformly follow accepted editorial style. As a technical requirement for constructing test items, this rule is absolute.

Sources for Learning Editorial Style

No one can be expected to have full recall of the accepted form of the thousands and thousands of various citations, only one or two of which may be needed in any particular test item. Fortunately, the well-informed item writer need not attempt to memorize correct citations because there are several very good references to which he or she can turn for assistance. (Even the Chief of Protocol for the United States, who stands by the President's side and whispers into his ear the titles and correct forms of address for various world leaders, does not have all of them memorized, but often relies upon a specially prepared book!) Item writers can be instantaneously informed on hundreds of citations, abbreviations, and correct forms of address with just one reference: *The Chicago Manual of Style*—or, for short, *The Chicago Manual.* Its full citation is given below, along with that of alternate, but also excellent, sources which may serve as references to editorial style:

- Achtert, W. S., & Gibaldi, J. (1985). *The MLA style manual.* New York: Modern Language Association of America.
- American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, D.C.: APA.
- Skilin, M. E., & Grey, R. M. (1974). *Words into type* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- *The Chicago manual of style* (14th ed.). (1993). Chicago: University of Chicago Press.

Dozens of other source books (many excellent ones) are also readily available in libraries and bookstores, and of course, more and more are becoming availabel on the Internet. Regardless of the particular title of choice, such a reference will aid the item writer in achieving this aspect of his or her technical task.

Editorial Style for Graphics

Correct formatting of graphs, tables, charts, illustrations, photographs, and other such graphic material is another important ingredient in good item construction. There are precise standards for editorial style of graphic

materials, and they too should be followed exactly. One should not presume that material in a graphic is "good enough" merely because the writer finds it appealing. Remember, consistency and uniformity of style help ensure that the graphic presentation is simple and clear.

As with other editorial stylistic concerns, a number of excellent sources exist to which one can turn for assistance in learning proper formatting of graphic material. Some of the better-known sources are as follows:

- Houp, K. W., & Pearsall, T. E. (1997). *Reporting technical information* (9th ed.). New York: Macmillan.
- *Illustrations for publication and projections* (ASA Y15. 1-1959). (1959). New York: American National Standards Institute. (Available from American National Standards Institute, Inc., 1430 Broadway, New York, NY 10018.)
- Strong, W. S. (1992). *A technical writer's handbook.* New York: Holt, Rinehart & Winston.
- Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

An additional note about the proper preparation of graphic material: Although selecting the appropriate type of graphic and formatting it correctly is exacting, even occasionally tedious work, it is nonetheless important, and the deliberate item writer will find it time well spent. Do not go lightly over this step when preparing test items.

Avoiding Biases in Language

Of course, good writers, whether preparing test items or in other writing, should avoid words that stereotype groups and language that may be offensive to individuals or outdated. This common-sense consideration is surprisingly difficult to keep in mind since most prejudices are subtle. Few writers deliberately use racist or offensive language in test items; but, subtleties of language can give rise to stereotypical or offensive language without the writer having intended offense. The following publications offer guidelines for reducing biases of ethnicity, age, gender, or disability:

- International Association of Business Communicators. (1982). *Without bias: A guidebook for nondiscriminatory communication* (2nd ed.). New York: Wiley.
- Maggio, R. (1987). *The nonsexist word finder: A dictionary of gender-free usage.* Phoenix, AZ: Oryx.

Sources for Specialized Editorial Circumstances

Several publications describe editorial style for specialized purposes. Publications that may be of use to writers of test items include the following.

For appropriate usage of mathematical expressions:

- American Institute of Physics. (1978). *Style manual* (3rd ed., rev.). New York: Author.

For the correct use of the metric system:

- National Bureau of Standards. (1979, December). *Guidelines for use of the modern metric system.* Dimensions/NSB, pp. 13-19.

For proper citation of legal procedures and proceedings:

- *A uniform system of citation* (13th ed.). (1981). Cambridge, MA: Harvard Law Review Association.

Although there are differences in many particular rules among various style manuals, item writers should uniformly follow one or the other of these guides to ensure clear, consistent presentation of printed test items.

USING DIRECTIONS CORRECTLY

Description of Directions

Directions for items are brief, single sentences instructing examinees how to respond after they have considered some relevant text or graphic, such as,

> **Read the poem below and answer questions 24 and 25.**
> **Use the map below to answer questions 12, 13, 14, and 15.**

Typically, items that do not have a text or graphic do not require individual directions.

Directions for items are not the same as directions for an entire test, which may specify procedural or logistical concerns, such as time limits or the use of a soft-leaded pencil. Overall test directions are often read to examinees by a test administrator. Usually, they are placed at the beginning of an entire test and apply uniformly to all test items in the entire examination. Directions for items, on the other hand, refer to only a single item or group of items that share some passage or graphic. And, when directions apply to more than a single item, they offer the further advantage of permitting stems for items to be simplified.

Clear, concise directions are an important part of well-constructed test items. Directions can specify the task for examinees by defining the activity required and focusing attention on relevant materials. Frequently, when responding to a test item, examinees must read a passage, study a diagram, or refer to a graphic. Unless the path from the text or graphic to the item's stem is brightly illuminated, examinees may not realize that a text or graphic accompanies a particular item.

Rules for Directions

The rules for writing directions are simple:

- Directions are only needed when a text or graphic accompanies an item.
- Directions should be clearly written and as concise as possible.
- Directions should appear above the relevant text or graphic and should refer the examinee's attention to the relevant item.

- Directions should appear in boldface type (for typeset tests or typewriters and word-processors with this capability).

These rules require elaboration for full understanding. Note that the first rule delimits directions for items to only those items that have an accompanying text or graphic. For all other items, directions are extraneous.

Table 5.1 Typical Directions for Items

A. For text inserts

1. Single paragraph:
   Read the paragraph below and answer question 19.
2. More than a paragraph:
   Read the passage below and answer questions 7 and 8.
3. Single sentence:
   Read the sentence below and answer question 57.
4. Poem:
   Read the poem below and answer questions 3, 4, and 5.

B. Graphic inserts

1. Map
   Use the map below to answer questions 1, *2,* and 3.
2. Diagram:
   Use the diagram below to answer question 51.
3. Table:
   Use the table below to answer question 9.
4. Graph:
   Use the graph below to answer questions 42 and 43.
5. Illustration (or picture):
   Use the illustration ("picture" for lower vocabulary levels) **below to answer question 22.**

Although unneeded directions may not be confusing to examinees, they add nothing and they tend to clutter the page.

A second rule for item directions—that the writing should be clear and concise—should be carefully noted. As much care should be taken in preparing the directions to an item as is spent in preparing the item itself. Table 5.1 presents a list of suggested wording for directions for a variety of situations. The suggested directions will apply to most situations. Item writers are encouraged to refer to this list and use the suggested wording when appropriate.

Another rule for item directions is that they should be located above the text or graphic to which they refer. It makes little sense to place text or graphic material first, where it may be read or skipped, and then instruct examinees in what they are to do. Misplacing directions can force some examinees to jump back and forth on the page between the text or graphic material, the directions, and the item itself. To facilitate the flow of information for examinees, the order should be: directions first, followed by the text or graphic material, and finally the item stem and response alternatives.

Illustrative Items 5.1 to 5.3 present examples of directions in correct editorial style. Illustrative Item 5.1 displays directions appropriately preceding text; 5.2 shows directions properly placed ahead of graphic material; and 5.3 displays directions correctly placed ahead of an item which has both graphic material and text.

Illustrative Item 5.1.

Read the passage below and answer the question that follows.

The driver and his mechanic were getting ready for the big race. The mechanic was making some last-minute adjustments to the engine. Bob, the driver, paced anxiously around the car. "Come on, Steve," he said. "Hurry up!"

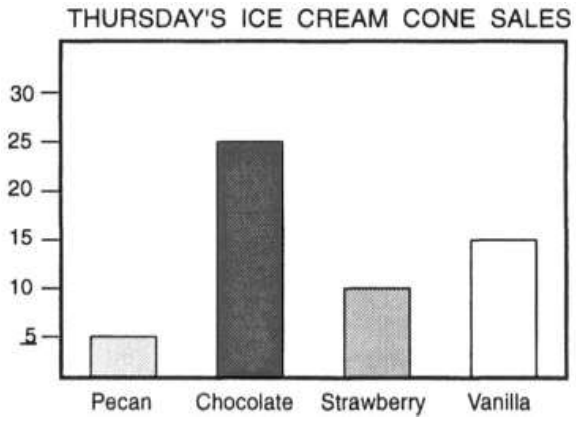
"Haven't you ever heard the saying Better safe than sorry'?" Steve asked. "Some things just can't be rushed."

What probably happened next?

A. Bob decided to drive a different car.
B. Bob fired Steve because he was too slow.
C. Steve finished his adjustments to the engine.
D. Bob jumped in the car and drove off.

Illustrative Item 5.2.

Use the graph below to answer the question that follows.

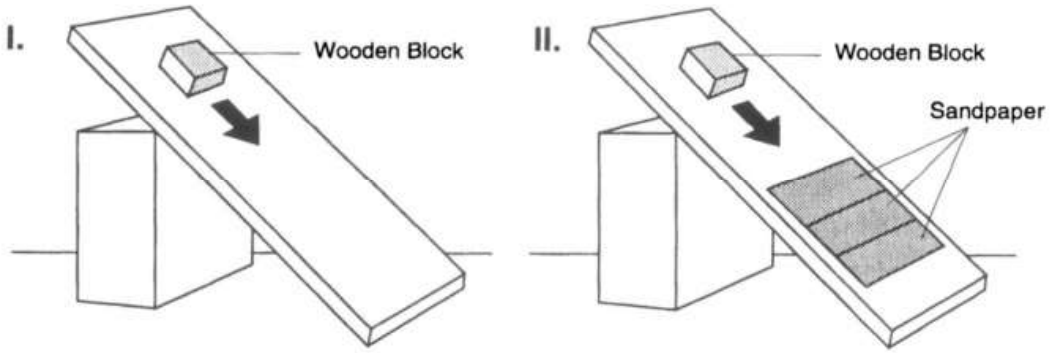


Which statement describes the information presented in the graph?

A. More butter pecan cones were sold than strawberry cones.
B. More vanilla cones were sold than strawberry cones.
C. More strawberry cones were sold than vanilla cones.
D. More vanilla cones were sold than chocolate cones.

Illustrative Item 5.3.

Use the diagrams and the paragraph below to answer the question that follows.



Lisa did an experiment in which she observed a small block of wood sliding down a long board. After trying this several times, Lisa taped sandpaper to the surface of the board. The block did not slide as far.

What stayed the same in this experiment?

A. size of the block
B. surface of the board
C. distance the block moved
D. time it took the block to come to a stop

Last, item directions should appear in boldface type. This draws attention to them so that there is less chance of their being missed by a nervous examinee. Also, it emphasizes their distinctness from the item stem. This printing requirement is reasonable, given the fact that today most typewriters, and nearly all word-processing programs and printers for computers, can accommodate boldface type.

Directions are easily accommodated in test items. Item writers should follow these rules for their correct and uniform use.

SPECIALIZED STYLE RULES FOR MULTIPLE-CHOICE ITEMS

A number of editorial and stylistic guidelines should be followed when preparing items in the multiple-choice format. These rules are simple to follow, and item writers should adhere to them consistently. They may be stated as a series of editorial rules.

Avoid Repetition in Options

Like stems, options should be as brief as clarity permits, and unnecessary repetitions should be avoided, as in Illustrative Item 5.4. In this case, the phrase "It is added" is unnecessarily repeated in every response alternative.

Illustrative Item 5.4.

When is paprika added to the omelet?

A. It is added before heating the skillet.
B. It is added before cutting the ham.
C. It is added while blending the ingredients
D. It is added after cooking the omelet.

should be

A. before heating the skillet
B. before cutting the ham
C. while blending the ingredients
D. after cooking the omelet

Omit Articles in Options

Articles may usually be omitted at the beginning of options that are not complete sentences, as shown in Illustrative Item 5.5.

Illustrative Item 5.5.

What part of Grendel did Beowulf remove?

A. a toe
B. the nose
C. the head
D. an arm

should be

A. toe
B. nose
C. head
D. arm

Handling Measurement Units

Whenever possible, units of measure should appear in the stem rather than in the response alternatives, as in Illustrative Items 5.6 and 5.7. In 5.6, it is unnecessary to repeat the word "inches" in every option, the wording in the stem clearly refers only to inches. In 5.7, it is obvious from the meaning of the stem that the options are all expressions of square centimeters.

Illustrative Item 5.6.

What is the length of the line?

A. 2 inches
B. 3 inches
C. 4 inches
D. 5 inches

should be

What is the length of the line in inches?

A. 2
B. 3
C. 4
D. 5

Illustrative Item 5.7.

What is the area in square centimeters of a sheet of wall paneling measuring 112 centimeters by 242 centimeters?

(A = l x w)

A. 354
B. 708
C. 13,552
D. 27,104

Note that this form is not possible when units are mixed, as in Illustrative Item 5.8. This is a case in which citing units is required for clarity.

Unless the objective to be tested requires the understanding of standard abbreviations for units of measure, the units should not be abbreviated in the stem or response alternatives:

2 mm
should be
2 millimeters

Illustrative Item 5.8.

What is the approximate length of this line segment?

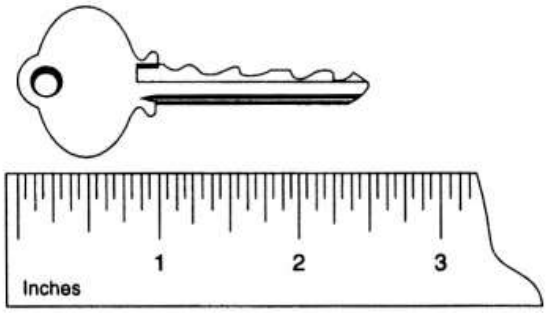
A. 2 inches
B. 3 inches
C. 2 centimeters
D. 3 centimeters

When the length of a line or object is the variable to be measured (as opposed to being approximated), the distances should be precise. This rule may be shown by comparing two presentations of the same item: one without exact specifications in the graphic for measurement, thereby making it necessary for examinees to approximate, and the same item improved by the addition of vertical rules, creating precision in measurement. Note these differences in the two versions of Illustrative Item 5.9.

Illustrative Item 5.9.

What is the length of the key to the nearest eighth of an inch?

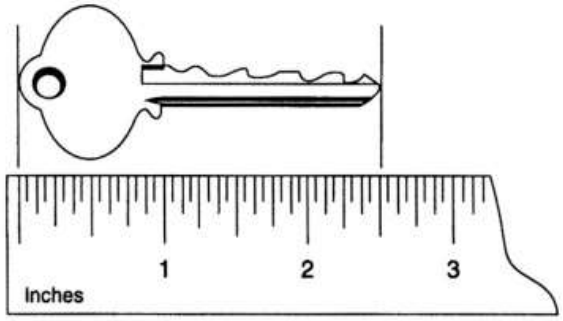
A. $2\frac{1}{8}$
B. $2\frac{1}{4}$
• C. $2\frac{1}{2}$
D. $2\frac{7}{8}$



should be

What is the length of the key to the nearest eighth of an inch?

A. $2\frac{1}{8}$
B. $2\frac{1}{4}$
• C. $2\frac{1}{2}$
D. $2\frac{7}{8}$



Use Third Person

Like stems, response alternatives should all be framed in the third person ("he," "she," or "it,") and particular care should be taken to avoid the second person ("you"). For example, notice in Illustrative Item 5.10 the rewording of the item's stem to the more accepted style.

Illustrative Item 5.10.

According to American Epileptic Society standards, what first aid should you perform when you witness someone having an epileptic seizure?

A. You should keep the person on his back.
B. You should turn the patient on his side.
C. You should give the person something to drink
D. You should put something between the person's teeth.

should be

According to American Epileptic Society standards, what first aid should be performed when someone has an epileptic seizure?

A. keep the person on his back
B. turn the patient on his side
C. give the person something to drink
D. put something between the person's teeth

Italics and Boldface

Italics are employed in accordance with conventional usage for foreign terms and for the names of books, plays, long poems, periodicals, paintings, films, musical works, ships, and like nouns. For example, Illustrative Item 5.11 correctly uses italics.

To consistently follow this rule, a manual of style, such as one of those cited earlier in this chapter, will probably be useful.

Illustrative Item 5.11.

Which is the best source for finding the 1980 population of Istanbul, Turkey?

A. *The World Almanac*
B. *Compton's Picture Encyclopedia*
C. *The Guinness Book of World Records*
D. *Webster's School and Office Dictionary*

Contrary to conventional usage in writing, however, boldface rather than italic type is used to designate words as words, as, for example, in Illustrative Items 5.12 and 5.13. Note particularly that the word is not boldfaced in the text, which should read as naturally as possible, but only in the item's stem.

Illustrative Item 5.12.

Read the paragraph below and answer the question that follows.

After several weeks at sea, the ship ran into bad weather. The crew had to work hard and their tempers grew mean. When the captain discovered the first mate sleeping on the job, he was furious. Then he learned that the mate had passed out from overwork. Everyone on board was eager to see fair skies again.

What does **overwork** mean in this paragraph?

A. hard work
B. lack of work
C. work above
D. too much work

Illustrative Item 5.13.

Which word means the same as **amiable** in this sentence?

Because Phillip was an amiable person, he was nice to everyone.

A. friendly
B. bashful
C. moody
D. mean

The rule for boldface type should also be applied if the words appear in the response alternatives, as in Illustrative Item 5.14. Again, note that these words are not boldfaced in the text.

Illustrative Item 5.14.

Read the paragraph below and answer the question that follows.

Ms. Theodore is trying to weasel out of paying a legitimate bill by saying we did not do the work on her car properly. I cannot believe that our respected legal system would let someone who has an honest debt evade paying it.

Which is a loaded word in this paragraph?

A. **weasel** because it makes Ms. Theodore seem sneaky
B. **properly** because it makes Ms. Theodore seem picky
C. **bill** because paying bills is unpleasant
D. **legitimate** because it is a formal word

Note that the rule extends to entire phrases, clauses, or sentences excerpted from a passage. To avoid confusion about what punctuation is included, quotation marks are not placed around the excerpt, as is shown in Illustrative Item 5.15

Illustrative Item 5.15.

Read the paragraph below and answer the question that follows.

1 I'd like to be like Nolan Ryan when I become an adult.
2 Bobby Grich was the last chance for the Twins from
3 Minneapolis, Minnesota and St. Paul, Minnesota
4 nevertheless he did his best. He gritted his teeth and
5 braced himself for what he felt was sure to be a fastball,
6 and no ordinary fastball the man on the mound was 29
7 years old Nolan Ryan the fastest pitcher in baseball.
8 One of his pitches were electronicly timed at 100.2
9 m.p.h. Well, back to Bobby Grich at the stadium. The
10 count was 3-2, and it was the bottom of the ninth. Ryan's
11 team ahead by a run. The fastball that Grich was so
12 sure of was a tantelizing change-up that glides over the
13 outside corner Ryan threw the third strike. Before he
14 could collect himself and swing at it, Ryan had done it
15 pitched his fourth no-hit game and ties the reocrd. The
16 newspaper in his town stated, "Ryan was the man of
17 the hour."

What is the correct form of **that glides over the outside corner Ryan threw the third strike** in lines 12 and 13?

A. that glides over the outside corner. Ryan throwed the third strike.
B. that glud over the outside corner. Ryan threw the third strike.
C. that glided over the outside corner. Ryan threw the third strike.
D. The form is already correct.

Boldface is also used instead of italics in a stem to emphasize negatives like not or least when the meaning of the stem would be reversed if the word were omitted. Examples are given in Illustrative Items 5.16 and 5.17. Notice in 5.16 and 5.17 that if "not" or "least" were overlooked by an anxious examinee the meaning of the sentence would be interpreted as the opposite of its true meaning. Boldface type makes it less likely that an examinee will miss these important words.

Illustrative Item 5.16.

Which animal is **not** a reptile?

A. snake
B. turtle
C. lizard
D. salamander

Illustrative Item 5.17.

For which essay would brainstorming probably be **least** effective when planning the first draft?

A. "How to be happy"
B. "Improving Cafeteria Menus"
C. "We Can All Conserve Energy"
D. "Normal Human Body Temperature"

In some situations, however, boldface is not required for least, as for example, in Illustrative Items 5.18 and 5.19. Here, the meaning of the sentence would not be reversed if the word not or least were omitted; rather, the sentence would be nonsensical.

Illustrative Item 5.18.

What is the least common denominator of these fractions?

$$\frac{1}{3}, \frac{1}{4}, \frac{1}{2}, \frac{4}{5}$$

A. 2
B. 4
C. 12
D. 60

Illustrative Item 5.19.

Which weighs the least?

A. hectogram
B. kilogram
C. megagram
D. gigagram

Mathematics items involving calculations are presented in imperative form and are indicated by boldface because they are considered directions, such as is shown in Illustrative Items 5.20 and 5.21.

Illustrative Item 5.20.

Add

29,647
+67,905

A. 87,552
B. 97,542
C. 97,552
D. 97,642

Illustrative Item 5.21.

Solve and simplify

$\frac{2}{7} + \frac{3}{7} =$

A. $\frac{1}{7}$
B. $\frac{5}{14}$
C. $\frac{5}{7}$
D. $\frac{6}{7}$

An alternative to this rule is to use this simple stem with all numerical operations: "Perform the following computation." Illustrative Item 5.22 presents this stem as an example.

Illustrative Item 5.22.

Perform the following computation

$$\begin{array}{r} 8{,}032 \\ \underline{-\ 3{,}938} \end{array}$$

A. 4,094
B. 4,194
C. 5,094
D. 11,970

What and Which

Often which and what are interchangeable, as in these stems: "What [Which] constitutional principles conflict in this case?" Or, "Which [What] is the topic sentence of this paragraph?" Still, to the fullest extent possible, a consistent use of what and which should be followed. Here are some guidelines: What usually implies an absolute answer; there are not degrees of correctness. For example: "What animal is this?" (Note that it can only be one animal.) Or, "What is the volume of this cylinder in cubic inches?" (Note that there is only one correct measurement.) This situation is also demonstrated in Illustrative Item 5.23, a circumstance in which there is one absolutely correct answer.

Illustrative Item 5.23.

What is the frequency in hertz of a sound wave with a velocity of 330 m/sec. and a wavelength of 2 meters?

A. 6.6
B. 16.5
C. 165
D. 660

On the other hand, which is frequently used as an abbreviated form of "which of the following," thus limiting the realm of choice to the options listed below the stem: "Which animal is a carnivore?" (Note there are other carnivores but none on this list.) Or, "Which cylinder has the greatest volume?" (Note that other cylinders that are not depicted have greater volumes.) Illustrative Items 5.24 and 5.25 show this distinction between which and what.

Illustrative Item 5.24.

Which nationality was most prevalent among the early European settlers of Manhattan Island and the Hudson River valley?

A. German
B. Scottish
C. Dutch
D. Swedish

Illustrative Item 5.25.

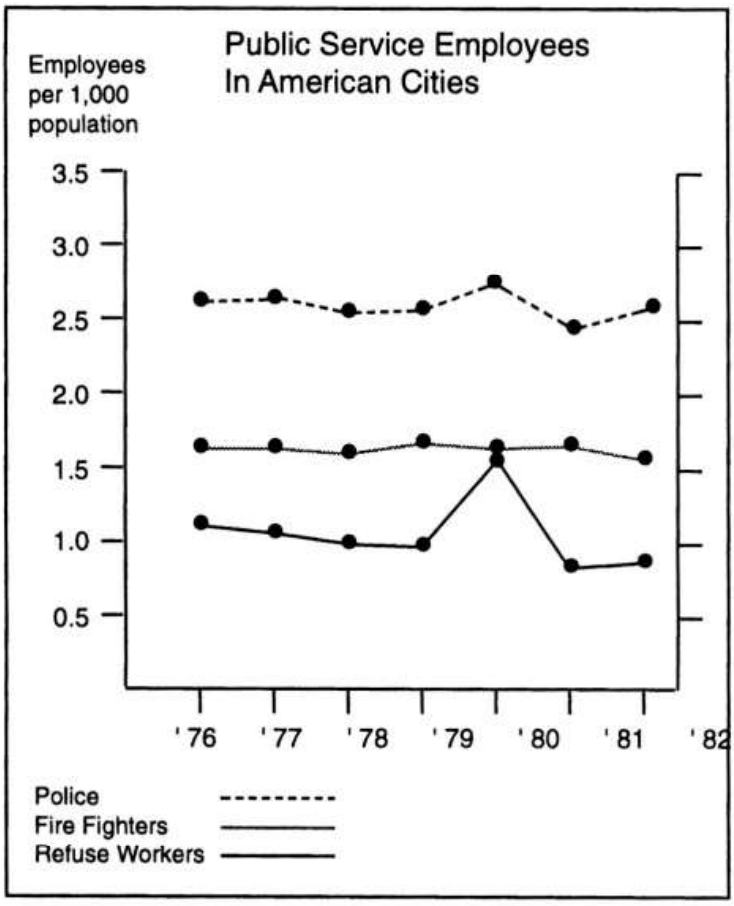
What is the supplement of an angle that measures 21°?

A. 69°
B. 111°
C. 159°
D. 179°

In a similar way, which may be preferable when referring to elements in a table, graph, diagram, illustration, or map, as is displayed in Illustrative Item 5.26.

Illustrative Item 5.26.

Use the graph below to answer question 77.



77. Which conclusion is supported by the graph?

A. American cities employed 26,000 police officers in 1976.

B. In 1980, American cities paid almost as much in salaries for refuse workers as they did for fire fighters.

C. The ratio of fire fighters to residents of American cities remained relatively stable from 1976 to 1982.

D. The decline in the number of refuse workers from 1976 to 1982 is the result of more efficient trash collection.

Ordering Response Alternatives

Ordinarily, response alternatives should appear in random order. However, numerals are arranged in ascending order, and dates are given chronologically, as in Illustrative Items 5.27 and 5.28.

Illustrative Item 5.27.

Add

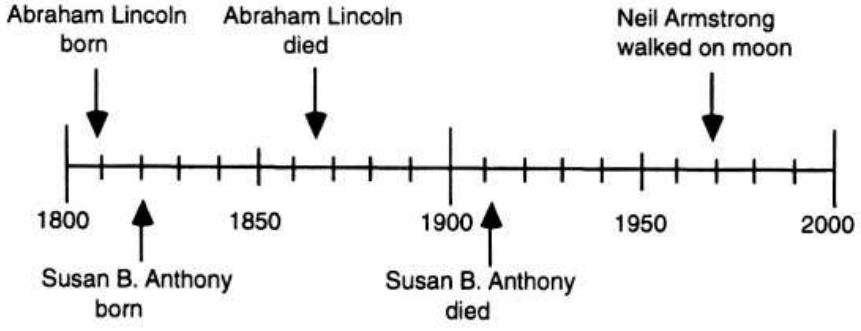
36
42
+17

A. 85
B. 86
C. 93
D. 95

Illustrative Item 5.28.

Use the time line below to answer the question that follows.

According to the time line, what was the approximate year of Abraham Lincoln's birth?



A. 1800
B. 1810
C. 1820
D. 1860

An exception should be made when such ordering provides clues to the correct response, as shown in Illustrative Item 5.29. In these cases, a random order is preferred.

Illustrative Item 5.29.

Which number is closest to zero?

A. $\frac{2}{3}$

B. $\frac{4}{7}$

C. $\frac{9}{64}$

D. $\frac{1}{2}$

Aligning Numerals

Numerals should always be aligned according to the following rules. Numerals without decimal points are aligned, or justified, on their right sides, as in Illustrative Item 5.30.

Illustrative Item 5.30.

Add

$$\begin{array}{r} 34 \\ 28 \\ 15 \\ +\ 86 \\ \hline \end{array}$$

A. 62

B. 163

C. 173

D. 1423

Decimals are aligned vertically on the decimal point, as is displayed in Illustrative Item 5.31.

Illustrative Item 5.31.

Which number has the same value as 5.031×10^3

A. 0.005031
B. 0.05031
C. 503.1
D. 5031

Dollar signs should also be aligned vertically, as is seen in Illustrative Item 5.32.

Illustrative Item 5.32.

A dress that ordinarily costs $82 is now offered at a 20% discount. If sales tax is 5%, what is the total cost of the dress on sale?

A. $ 6.51
B. $65.60
C. $68.88
D. $69.70

With the exception of money values, if decimals less than 1 are grouped with values greater than 1, a 0 is used to fill the one's place, as shown in Illustrative Item 5.33.

Illustrative Item 5.33.

The density of a cube of ice is $0.92\ g/cm^3$ and its mass is 13.8 g. What is the volume of the cube in cm^3?

A. 0.15
B. 12.6
C. 15
D. 126

Numerals of two digits or less may be spelled out in most cases. However, mathematics items should employ Arabic numerals for all numbers.

TYPE CHARACTERISTICS AND PAGE LAYOUT

Importance of Appearance

Considering the appearance and arrangement on the page of test items is a part of item construction. An attractive appearance will facilitate communication with examinees who may otherwise be distracted by sloppy page layout, difficult-to-read type, or poor-quality type. Additionally, insuring an attractive appearance for the test items that one has so carefully crafted should be a final step in a professionally produced product.

Formatting test items attractively can be readily done with today's technology. Many modern tests, even those that are "home made," are constructed with the aid of sophisticated word-processing and page-layout programs on personal computers. This is a boon to item developers because the increased sophistication in formatting capability permits great flexibility in many production features, such as selection of the typeface, and type size and the arrangement of items on the page.

This section will describe some suggestions for the appearance features of test items. This is not to imply that these guidelines are absolute; rather, they will provide some information which may be useful. When considering the typeface, size, and page layout, each item writer should consider the needs of a particular situation as well as the resources which he or she has available.

Suggested Typefaces for Items

Typeface is the style of type in which an item is printed. Any of several typefaces may be suited for tests, but for uniformity, it is recommended that item writers select one of three typefaces: Helvetica, Palatino, or Times Roman. These typefaces are standards throughout the printing industry, and any major-brand word-processor or modern typewriter will accommodate all of them. They are easy-to-read typefaces whose appearance will be

instantly familiar to examinees. If one is preparing tests on a typewriter or other machine that does not accommodate these typefaces, then Courier type (the standard for most early typewriters) should be used.

When the choice of a typeface is made deliberately and artfully, it is doubtful that examinees will pay a second thought to the particular typeface chosen. This is as it should be, because if examinees did notice the type, it would almost certainly be a distraction. Attending to the detail of selecting a pleasing typeface and size, however, is a task of the skilled item writer. Except when otherwise noted, all of the illustrative test items in this book are printed in Helvetica type.

Helvetica type is a sans serif type. A serif is the little end mark, or curly-cue to letters. Letters that do not have a serif are called sans serif, or sometimes, block letters. The reader can easily spot Helvetica type by examining most of the illustrative items in this book. (There are a few items presented later in this chapter that are not printed in Helvetica; rather, they are printed in other typefaces for illustration.)

Alternatively, Palatino and Times Roman are serif typefaces. Both are attractive, and either can be used. The text in this book is printed in Times Roman. In some circumstances, these typefaces have advantages over Helvetica. For example, when Helvetica is printed in a large size, its block appearance can seem clumsy. In a large type size, the rounding curl of a serif adds a visual follow-through for the eye. Hence, it is recommended that when tests are to be printed in a large type size—as, for example, tests that are intended for use with very young children—Palatino or Times Roman should be used.

Suggested Type Size for Items

Type size is the physical dimension, from bottom to top, of the type. Except for specialized printing (signs and so forth), type size can range from one to 100 points. One-point type is so small as to be illegible (actually a type size this small is never used), and 100-point type is about one-and-a-half inches tall. But, most type lies somewhere between six-point and 48-point. The text of this book is printed in 12-point type. Often typefaces are measured in picas rather than point size. The conversion between the two is easily made: one pica equals 12 points.

For attractive test items, it is recommended that 10-point type be used for tests intended for adults or for students in grade 10 or above. For tests

to be used with younger students who are still gaining proficiency in the visual acuity necessary for reading (from about grades 7 to 9), 12-point size is recommended. And, for tests that are intended to be used with very young children, who are just developing the visual acuity needed to read (from kindergarten to grade 6), 14-point type may be most easily recognized.

Table 5.2 presents a recommended typeface and type size for printing tests intended for various grade and age levels.

The reader can compare the difference in these typefaces and type sizes by scrutinizing Illustrative Items 5.34 to 5.36.

Placing Passage and Graphic Dependent Items Together

Another important aspect of appearance is the proper placement of items on the page. This aspect of appearance facilitates examinees' approaching the items in the clearest manner possible. The guidelines for this consideration are straightforward. In most cases, even when an item contains a long reading passage or a large graphic, the item's stem and response alternatives should appear on the same page as the text or graphic. In rare instances when an item requires more space than a single page, such as one that incorporates lengthy text or a very large map, the text or graphic can appear on one page and the item's stem and option on the next. In these cases, the text or graphic should appear on the verso (the left-hand or even-

Table 5.2 Recommended Typeface and Type Size for Printing Tests Intended for Various Grade and Age Levels

Grade Level or Age at Which Test is Intended for Use	Typeface	Type Size
K to Grade 6	Palatino	14
Grades 7 to 9	Helvetica	12
Grades 10 and above	Helvetica	10
Adult	Helvetica	10

Illustrative Item 5.34.

Read the paragraph below and answer the question that follows.

Not long ago a scientist quietly watched the daily habits of a pair of robins and their young. At three o′clock in the morning, the hard-working parents began feeding the babies. After their breakfast, the scientist saw the young mother bird wash and comb her young. Drawing each tiny feather through her bill, she removed every speck of dirt with her moist tongue. This was done every day until the young robins were old enough to clean themselves.

When did the mother robin wash the babies?

A. at three o′clock
• B. after breakfast
C. after they flew
D. in the evening

Illustrative Item 5.35.

Read the paragraph below and answer the question that follows.

Not long ago a scientist quietly watched the daily habits of a pair of robins and their young. At three o'clock in the morning, the hard-working parents began feeding the babies. After their breakfast, the scientist saw the young mother bird wash and comb her young. Drawing each tiny feather through her bill, she removed every speck of dirt with her moist tongue. This was done every day until the young robins were old enough to clean themselves.

When did the mother robin wash the babies?

A. at three o'clock
• B. after breakfast
C. after they flew
D. in the evening

Illustrative Item 5.36.

Read the paragraph below and answer the question that follows.

Not long ago a scientist quietly watched the daily habits of a pair of robins and their young. At three o'clock in the morning, the hard-working parents began feeding the babies. After their breakfast, the scientist saw the young mother bird wash and comb her young. Drawing each tiny feather through her bill, she removed every speck of dirt with her moist tongue. This was done every day until the young robins were old enough to clean themselves.When did the mother robin wash the babies?

A. at three o'clock
B. after breakfast
C. after they flew
D. in the evening

Figure 5.1. Examples of how to place items with text or graphics in a test.

Reading/Language Arts Page 7

Read the paragraph below and answer question 17.

Have you ever looked up to see the antlike figures of men working on a huge building? Did it make your head spin and your palms sweat? Well, the Mohawk Indians think nothing of walking across a narrow steel beam that rises dozens of stories in the air. Their skill in building giant skyscrapers and bridges is known far and wide. The Mohawks are among the finest steelworkers in the world.

17. At what other job might a Mohawk Indian steelworker excel?

    A. architect
    B. tightrope walker
    C. ant farmer
    D. bridge operator

Read the paragraph below and answer question 18.

I know where the crystal flash is hidden. I was the one who climbed up the ladder and hid it. But I would never dare tell anyone. Hush! Here comes the princess.

18. What is the point of view in this paragraph?

    A. first person
    B. second person
    C. third person
    D. cannot be determined

Read the paragraph below and answer question 19.

It has been proven by a team of scientists that seals have an acute sense of hearing. Seals were trained to recognize different sounds. For instance, the sound of one bell meant food, and two bells meant no food. If the seals responded incorrectly, they were given a light tap. It was discovered that seals are very good learners. They were easily taught the difference between the sounds.

19. Which word is a synonym for **acute** as it is used in this paragraph?

    A. sharp
    B. loud
    C. weak
    D. unusual

20. Which sentence states a fact rather than an opinion?

    A. Blue jays are greedy, noisy birds.
    B. Blue jays are prettier than cardinals.
    C. Blue jays are a type of bird.
    D. Most people dislike blue jays.

Science

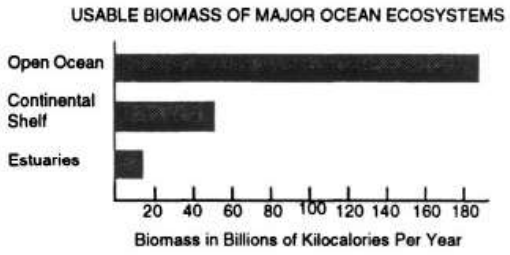
Use the table below to answer question 42.

Type of Light	Sun	Green	Blue	Orange
Average plant height at start of experiment	23 mm	22 mm	22 mm	21 mm
Average plant height one month later	51 mm	24 mm	41 mm	43 mm

42. The use of blue light resulted in a height increase of 95%. Which percentage correctly describes the increase under orange light?

A. 47%
B. 56%
C. 70%
D. 105%

Use the graph below to answer question 17.



17. According to the graph, which statement is true?

A. The open ocean has the greatest usable biomass.
B. The estuaries have the greatest usable biomass.
C. The continental shelf has the great est usable biomass.
D. None of the areas has more usable biomass than the others.

74. An automatic drip coffeemaker has a power rating of 1000 watts, and a stereo has a power rating of 80 watts. How many kilowatt hours are consumed when the two appliances are used continuously for 3 hours?

A. 2.40
B. 3.00
C. 3.24
D. 3240

75. Why does tire pressure increase when air is pumped into an automobile tire?

A. The tire expands, increasing the inside surface area.
B. More air particles exert more force while the inside surface remains almost constant.
C. The force exerted for a given area of inside surface decreases when more particles are added inside.
D. The force exerted y the particles inside the tire is reduced because the particles collide more with each other.

76. Which reaction will produce the greatest amount of energy?

A. $^{235}_{92}U + {}^{1}_{0}n \rightarrow {}^{141}_{5}Ba + {}^{92}_{36}Kr + 3{}^{1}_{0}n + \text{energy}$
B. $4{}^{1}_{1}H \rightarrow {}^{4}_{2}He + 2 + {}^{0}_{1}e + \text{energy}$
C. $CH_{4(g)} + 2O_{2(g)} \rightarrow CO_{2(g)} + 2H_2O_{(g)} + \text{energy}$
D. $H_2O_{(g)} \rightarrow H_2O_1 + \text{energy}$

77. How many kilowatt-hours are used by a 4-watt electric wall clock in 30 days?

A. 0.096
B. 0.12
C. 2.88
D. 5.0

Mathematics

92. A right circular cone has a height of 12 inches and a base with a radius of 5 inches. What is the volume in cubic inches of the cone?

Formula:
$v = \frac{1}{3} h \pi r^2$

A. 20π

B. 100π

C. $\frac{437\pi}{3}$

D. 300π

93. What is the perimeter in centimeters of an isosceles triangle with sides of 11 centimeters and a base of 6 centimeters?

A. 17
B. 28
C. 33
D. 66

94. A right triangle has legs that are 6 inches and 8 inches. What is the length in inches of the hypotenuse?

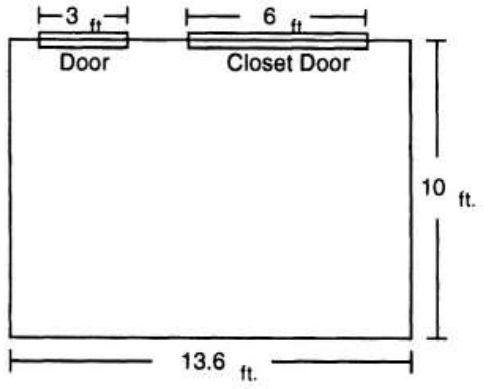
A. $3\sqrt{5}$

B. $3\sqrt{13}$

C. 10

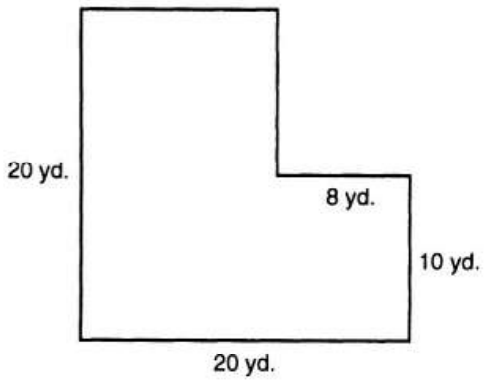
D. 15

95. Excluding the door and closet, how many feet of baseboard molding is needed for the perimeter of this room?



A. 14.6
B. 23.6
C. 38.2
D. 47.2

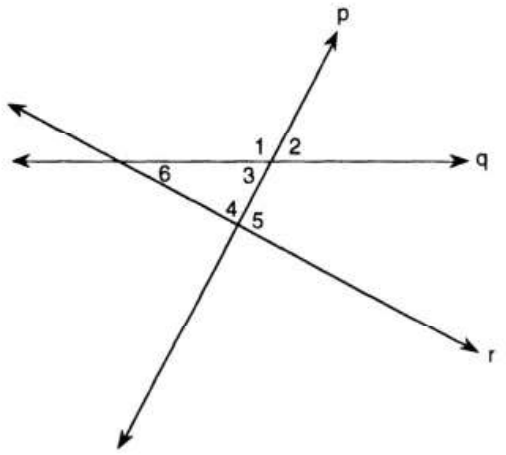
96. How many square yards of sod would be the minimum necessary to cover this yard?



A. 160
B. 240
C. 320
D. 400

Mathematics

Use the diagram below to answer questions 88 and 89.



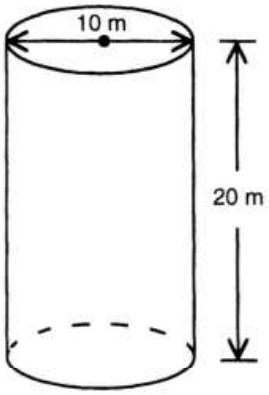
88. Which are vertical angles?

A. 1 and 2
B. 2 and 3
C. 3 and 5
D. 3 and 6

89. Which are supplementary angles?

A. 1 and 3
B. 2 and 3
C. 3 and 4
D. 4 and 6

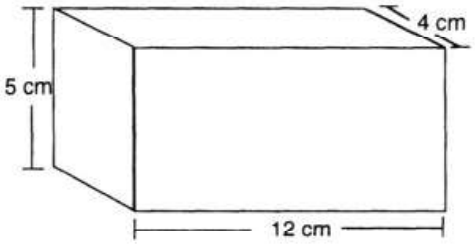
90. What is the volume in cubic meters of this cylinder?



Formula:
$v = \pi r^2 h$

A. 500π
B. 1000π
C. 2000π
D. 2500π

91. What is the surface area in square centimeters of this wooden block?



A. 80
B. 128
C. 232
D. 256

numbered) page, and the item's stem and options on the recto (right-hand or odd-numbered) page. Under no circumstances should examinees be forced to turn pages to refer to all parts of an item.

With these considerations addressed, when a test is finally produced it will reflect the quality of construction used for the items. Not only will this include the technical issues described earlier in this book, but it will also reflect the appearance. Figure 5.1—presented on the preceeding pages—exemplify careful attention to these concerns.

CONCLUSION

This chapter focused on design and formatting characteristics of test items in the multiple-choice format. It began with a thorough discussion of the multiple-choice format for items, citing advantages and criticisms. This was followed by a explanation of editorial guidelines and a description of specialized rules for formatting multiple-choice items. Finally, the appearance of finished page proofs for items was addressed, including appropriate type specifications and appealing page layouts.

By adhering to these editorial and style rules for formatting multiple-choice items, the item writer can produce items that will be uniform in format and consistent in style, and which can be presented to examinees in appealing fashion. Each of these is important considerations in the professional preparation of test items. Of course, many of the rules and guidelines presented can be applied to test items in other formats. How and when they should be applied will be described in the following chapter.

Chapter 6

Style, Editorial, and Publication Guidelines for Items in Constructed-Response/Performance Formats

INTRODUCTION

In the early 1990s, multiple-choice testing was itself put to the test. School achievement seemingly had not improved despite widespread use of large-scale testing programs. Critics opined that the tests were integral to the problem. Curriculum designers had invented new methods for delivering instruction in basic skills, and other, older teaching methods were rediscovered and began to make inroads into the classroom. The tests, it was widely argued, did not "fit" with these new curriculum efforts. Educators, both individually and collectively as they may have organized themselves into myriad associations and groups, called for massive, radical change, especially in the assessment arena. The public, too, was tiring of hearing new education slogans and seeing more reports documenting the lack of progress and the changing demographics of the student population. They were apparently ready to accept new assessment procedures. Even the U.S. Congress requested information on the debate (cf. "Testing in American Schools: Asking the Right Questions," U.S. Congress, Office of Technology Assessment, 1992). Clearly, the *zeitgeist* argued for something new to

come along and save the day from the dreaded multiple-choice test item format.

What, specifically, was to be the new item formats for assessment was unclear, so long as it was away from multiple-choice test items. A variety of approaches were proffered under an array of names: performance assessment, constructed-response, higher-order assessment, portfolio assessment, authentic assessment, and many others. Tests comprised of items and exercises in any and all of these alternative assessment formats were tried.

While obvious and substantial differences exist among the various alternative assessment strategies—and the item or exercise formats that they may assume— no commonly accepted set of terms for their description has emerged. One feature about them, however, seems ubiquitous: their eshewance of A, B, C, D-type response.

Formats for items in these alternative strategies usually require an examinee to either construct a response to a particular stimuli (e.g., write in a short answer or essay, devise a solution to a problem, describe how something may be constructed) or to perform in some way (e.g., artistic expression, musical or thespian performance). Hence, we will refer to these formats generally as "constructed-response" or "performance-response." Further, since the purpose in this chapter is to provide direction for formatting items and exercises in the newer approaches for style, editorial aspects, and other publication guidelines, the terms are used interchangeably; again, recognizing that true differences exist in the various assessment strategies.

From one technical perspective, the term "item" does not adequately cover the myriad formats that stimuli in the alternative approaches can assume. "Exercises" may be a more generally appropriate descriptor; but, since almost everything described about test items also applies to the exercises, for practical purposes, the terms can be used interchangeably.

The reader is asked to remember that this chapter provides only a brief introduction to these alternative assessment strategies, and focuses on style, editorial, and publication guidelines. A huge and ever-burgeoning body of literature and information is available about these approaches and the interested reader is directed to consult additional sources for information about measurement characteristics, developing scoring schemes, and other information for the alternative strategies. Hambleton (1996) offers an excellent review and discussion of advances in assessment models, meth-

ods, and practices. Readers are particularly directed to Hambleton's work. Also, Hymes (1991) provides a resource for information about the alternative approaches. And, many colleges or universities that offer teacher preparation programs have vast amounts of information about the alternative approaches to assessment. For example, UCLA's CRESST (Center for Research on Evaluation, Standards and Student Testing), located within the College of Education, produces a wealth of information about alternative assessment and makes it available in a variety of formats (written publications, reference lists, CD-ROMs, etc.).

TYPES OF CONSTRUCTED-RESPONSE AND PERFORMANCE

As mentioned, no uniform terminology exists to describe particular formats or types of items and exercises in the constructed-response or performance categories. Hence, what one may consider a performance, another may view as a simulation, or something else. Nonetheless, it is useful to consider formats which the alternative approaches may assume. Table 6.1, below, lists a number of options for formats appropriate to the alternative assessment strategies.

Bennett and Ward, in their edited volume *Construction Versus Choice in Cognitive Measurement: Issues in Constructed-Response Performance Testing and Portfolio Assessment* (1993), offer a rich resource for exploring the definitions of various constructed-response and performance formats, as well as discussion of the distinctions between them and traditional selected-response formats.

From the various formats cited in Table 6.1 one can readily deduce a number of characteristics that are common to most constructed-response and performance formats. First and perhaps most obvious, these alternative formats require an examinee to supply, develop, perform, or create something. And, typically, these tasks attempt to be more engaging to the examinee than traditional multiple-choice items. Often, they employ real-world problems that people of a comparable age and peer status may encounter in daily life, such as asking school-age children to calculate change from a grocery store purchase, or for high schoolers, to complete a driver's license application or examine an insurance policy.

Alternative assessments are generally scored by comparing and contrasting the examinee's response to some developed criteria. Further, the criteria typically have a range of values corresponding to degrees of

Table 6.1 Illustative Constructed-Response Item Formats

Cloze Procedure	Observations
Demonstrations	Oral Reports
Discussions	Performances
Essays	Portfolios
Exhibitions	Projects
Experiments	Research Papers
Fill-in-the-Blank	Reviews
Grid-in Response	Self/Peer Assessments
Interactive Video	Short Answer
Interviews	Writing Samples

completeness in satisfying the criteria. Sometimes, especially when the assessment requests a written product (i.e., essay), the criteria are elucidated in lengthy descriptions called "rubrics." For a discussion of developing and evaluating scoring criteria for alternative assessments, the reader is referred to Herman, Aschbacker, and Winters (1992) and to Linn, Baker, and Dunbar (1991).

VALIDITY IN CONSTRUCTED-RESPONSE

When considering validity for alternative assessment formats, one is cautioned to not give it little weight, or even ignore it altogether, as sometimes happens when individuals are too eager to throw out multiple-choice items. Hambleton (1996) cautions, "It should be emphasized that the new items formats, although more attractive in some respects than the common multiple-choice format, are still required to meet psychometric standards of excellence" (p. 905). Regretfully, the lack of attention to validity, and to other psychometric considerations of good measurement, is commonplace, according to Hambleton. Many other researchers back up his point.

When individual schools, public school districts and even some state education agencies learned of the alternative assessment approaches, they were eager to jump on the do-away-with-multiple-choice-format band-wagon. The wisdom of such a hasty decision is now being questioned. In one

famous example, in 1991, the state of Vermont initiated a complex and comprehensive system of capturing a variety of student products, and then cataloging them and scoring them. The Vermont statewide assessment initiative called the collection "portfolios." Vermont garnered national attention with its program, but it had little concrete evidence that its portfolios provided useful information. After debate, a format evaluation study of the program was conducted by RAND (cf. Koretz, Klein, McCaffrey, & Stecher, 1993) which concluded that the program scoring had very low reliably and its validity was suspect or unknown at best.

Recently, others have investigated performance measures more generally and have come to similar conclusions. Mehrens (1992) suggests that such measures generally lack reproducibility and therefore cannot be considered reliable assessments. Haertel (1990) reached the same conclusion when he stated that research has failed to demonstrate reliability, a hallmark of good assessment. Beck (1991) argues that lacking the fundamental measurement characteristics of reliability and validity, performance assessments are not "socially, politically, or legally viable" and should not be implemented on a large-scale.

Perhaps the most regrettable aspect of the all-too-common-inattention-to-validity phenomenon among zealous advocates of the alternative formats is their misunderstanding of the role of assessment. Assessment is meant to uncover information about mental constructs as manifested in achievement, ability, or aptitude. It is not a cause *celebre*. One good rule-of-thumb about validity when constructing exercises in alternative formats is to simply ask whether it can be readily agreed upon by appropriate experts what construct is being tapped. In other words, ask the question, What are we measuring by this item or exercise? For evidence for validity in interpretations to the exercise, the construct should be clear, precise, and reasonably distinct from other constructs. To imagine that "higher order thinking skills" is a single construct, or that "problem solving" is a uniform aspect of achievement, aptitude, or ability, is to misunderstand assessment and even displays an ignorance about it. Constructing test items and exercises, whether in the traditional or the alternative formats, is too important to not address central concerns.

Table 6.2, adopted from work by Herman, Aschbacker, and Winters (1992), presents considerations for constructing items and exercises in alternative assessment. Following these guides should help one produce meritorious exercises in alternative formats.

Table 6.2 Considerations for Alternative Assessments

- What outcome(s) are intended for the assessment?
- What are the eligible content/topics?
- What is the nature and format of questions to be posed to students? What is the audience for the response?
- Is it group or individual work? If group work, what roles are to be filled?
- What options/choices are allowed? What are the choices in response mode? What will they include, for example portfolios? Who makes the choices—the teacher or students or both?
- What materials/equipment/resources will be available to students? Are there any specifications?
- What directions will be given to students?
- What administrative constraints are there? How much time is allowed? What is the order of tasks? How will student questions be answered? What help will be allowed?
- What scoring scheme and procedures will be used?

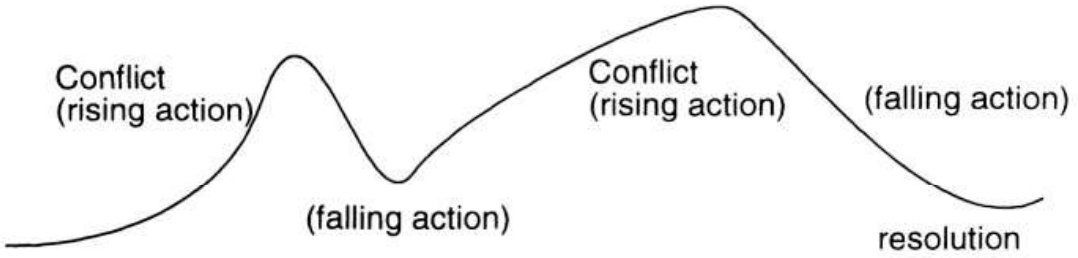
PRECISION IN WORDING, AGAIN

The preceding chapter emphasized the importance of good writing when constructing items in the multiple-choice format. The rationale for clear wording used for that format—good communication and clarity of scientific thought—also applies to constructed-response and performance assessments. Only here the notion of good writing is more easily edited because most constructed-response and performance exercises are longer than one sentence. Some consume much space when giving the pertinent information and directions. However, like all writing, whether lengthy or brief, it is easy to lose clarity. Below, are some examples of poorly constructed items and exercises in the alternative formats.

To begin, consider Illustrative Item 6.1, which gives an example of unnecessary and awkward wording for an item in the constructed-response format. The item was selected from a state assessment program, and it is intended for use with twelfth-graders. It was developed by a major test publisher.

Illustrative Item 6.1

In the story "Sled," there are actually two crises. Study the diagram below. It represents the plot elements of the story. Then, read the sentences that follow. Each sentence indicates a particular point in "Sled." Decide into which of the six plot elements each sentence falls. Before each sentence, write the appropriate letter in the blank?



_____ With his hands he grasped her head and titled it so that the light fell diorectly on the cut.

_____ He saw her take a package from the man and come out of the store.

This item is confusing to students—even seniors in high school, the grade-level for which it was intended—for a number of reasons. First, the item stem (or scenario) statement does not provide clear directions. In fact, it includes extraneous information and omits necessary information. As was demonstrated repeatedly, meritorious test items should not be "tricks" designed to catchup a student in an intentional web. Willful deceit is no substitute for well-crafted test items. Additionally, the item is confusing because it is not formatted in a manner that facilitates clarity; rather, it seems

to be rather sloppily thrown together. The sections of the item do not flow smoothly from one to the next; instead, the examinee is left to figure it out.

Illustrative Item 6.2—an item from the New Standards Project, and one that has had widespread, national use—allows us to examine style and editorial characteristics for layout of items. Note that this item is marginally clear by providing minimal instruction to the examinee. It presents the scenario first, followed by a list of conditions that pertain to the scenario, and finally the directions. The title is also only of limited value to examinees. Editorial features for the item could be improved by providing the directions first and then the scenario. A graphic (such as a picture of some children playing checkers) would make the item more engaging, but it is not absolutely necessary to the content of the activity.

Additionally, consider the directions provided to examinees for Illustrative Item 6.2. They are vague and indeterminate, and regretfully reflect the fact that the developers of the exercise are seemingly unclear about what mental construct is being assessed, a fatal flaw discussed above in validity considerations. Providing clearly focused directions is integral to constructing meritorious items in alternative formats.

A note of elaboration on graphics may be useful, too. Increasingly, items in all formats, whether traditional or alternative, employ graphics. This is not a bad trend, but like all aspects of item development, they must be considered carefully. The essential point to bear in mind here is that the graphic should be related to the stimuli in some inherent way. Merely adding a picture to an item or exercise for no reason is not appropriate.

Illustrative Item 6.2

Game of Checkers

Marc, Anna, Julia, and Daniel decided to have a checkers tournament at school. They want to be sure each of them gets a chance to play each of the others one time. They ask you to make a schedule for the tournament. Here is the information you need to make a plan that works.

- They want to finish the tournament in one week. They can play from Monday through Friday.
- The will play only at lunchtime. There is enough time during lunch period to play one game of checkers.
- The students have two checker sets, so two games can be going on at once.
- Marc can't play checkers on the days he is a lunch helper (Mondays and Wednesdays).
- Each player must play every other player once.

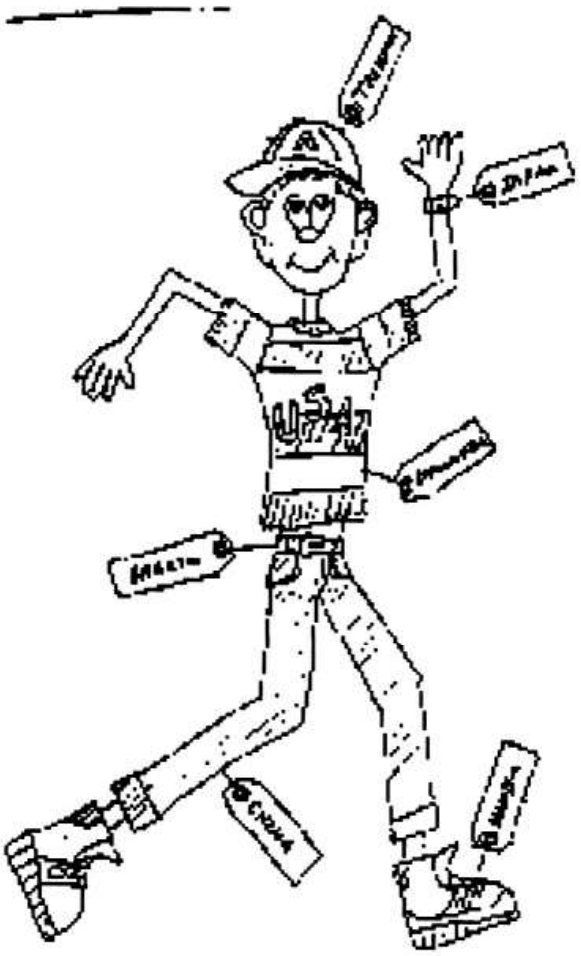
DIRECTIONS:

Make a schedule for the tournament, using both written form and graphic form. It should be creative. Your schedule should make it easy for everyone to see who plays whom each day. Then prepare a series of charts, one for each day of the tournament, showing who won among the matches for that day as well as the odds of the forthcoming matches. Finally, devise a set of handicapping rules that would give the losing players more chances to win against the winning players.

Item developers today have easy access to sophisticated graphic-development-and-enhancement computer programs. There is no excuse for poor art work, such as is shown in Illustrative Item 6.3, an item in a national testing program developed by a major test publisher and whose poor quality is reproduced here exactly. Regretfully, this item painfully demonstrates the lack of care given to graphics by lazy or uninformed item developers.

Illustrative Item 6.3

Write a caption for the cartoon below.



(Scoring key: for 1 point the caption should have 1 reasonable sentence relating to the cartoon; for 2 points the caption should relate to the theme of global interdependence; for 3 points the caption should include 1 and 2 above plus be humorous.)

Finally, we come to a well designed and carefully thought-out item in the alternative format, Illustrative Item 6.4. Notice first that the construct for assessment is clearly identifiable, achievement in basic mathematical operations. Also noteworthy is that the item has multiple responses but not an infinite number of correct answers. This desirable feature makes devising scoring criteria a simple, straightforward task.

Importantly, the directions provided to examinees in the item or exercise are clear and thorough. And, finally, it is attractively presented, with the directions given first and the exercise following. This items is a good model for item developers to follow.

Illustrative Item 6.4

Directions: In this mathematics problem you can use addition, subtraction, multiplication, and division. And you can use each more than once. You are to find as many correct results from the data in the table as you can. Here's how to do it. First, select a number from Column I. Next, chose an operation from Column II. Then find a number in Column III that will give you a correct result from among the choices in Column IV. Find as many answers as you can.

I	II	III		IV
2	+	1	=	1
3	−	2	=	3
4	×	3	=	5
6	÷	4	=	6

CONCLUSION

This chapter focused on items and exercises in formats that are considered alternative to traditional selected-response items. It presents information about the items, identifying them and paying special attention to valid interpretation from scores that they can yield. Attention was given to editorial style but, because there is an unlimited range of formats for such items, no exhaustive set of editorial guidelines is offered. Instead, the reader is referred to some basic rules, mentioned above, such as attention to validity, good writing, and attractive presentation.

The information in this chapter, taken together with that offered in the preceding chapter and the following one, should give writers a store of information about presenting items in both traditional and alternative formats.