Constructing Test Items:

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

Second Edition

Chapter 2

Definition, Purpose, and Characteristics of Items

INTRODUCTION

An elementary rule for any writer is to know thoroughly the subject before beginning to write. This is the reason reporters ask questions, researchers investigate hypotheses, and novelists ponder their protagonists, all before they put pencil to paper. So too must writers of effective test items master their subject. Learning about test items means comprehending what test items are, understanding their purpose, and becoming familiar with their characteristics. Knowing the definition of a test item is the first step toward comprehending it. The definition forms the backbone upon which the purposes and characteristics of test items rest.

Learning the purposes for test items is another essential ingredient of comprehension. Test items derive their purpose from psychology; accordingly, some background in behavioral theory is necessary to appreciate their purpose as tools for measurement.

Becoming familiar with the characteristics of test items is also important. Characteristics of test items include the various formats that they may take, their essential components, and necessary assumptions underlying their use. Characteristics also define the item by delimiting the type of scores that an item can yield, permitting or disallowing its use in particular situations, and dictating practical considerations, such as the time examinees need to respond.

Some characteristics of items apply generally to all of them, while others pertain only to certain types. For instance, all test items present a stimulus and prescribe a response, but the nature of the stimulus and response of a true-false item is unique to its type. Similarly, the item forms of multiple-choice, matching, sentence-completion, and cloze-procedure each have their unique characteristics. Constructed-response formats are likewise bound by their own set of characteristics.

Another dimension of test items, intertwined with their characteristic features and qualities, is their inherent theoretical basis. It is important for conceptual consistency that test items stem from a theory of measurement and that their particular characteristics complement the theoretical framework. If a test item is to be used properly for psychological measurement, one must understand the relevant theoretical assumptions and give them ample consideration throughout the process of developing the item.

The reason it is important to thoroughly understand the definition, purpose, and characteristics of test items before writing them may seem obvious, but it is so significant that it warrants being stated. By knowing the definition, purpose, and characteristics of test items, one will have at hand a great deal of information about a particular test item, its construction, function, and probable effectiveness. Informed and diligent test-item writers are more likely to produce items of quality—that is, test items that meet criteria for good items—than may be yielded with a haphazard approach to item construction by well-intentioned but uninformed persons. An uninformed approach to constructing test items may lead to a set of questions that can be neatly printed on a page and give the appearance of a test instrument, but will more likely result in gross errors in measurement.

Finally, when discussing the definition, purpose, and characteristics of test items, a standard, professional terminology should be used. Currently, no common vocabulary is followed by most item writers There is a need to standardize terms, descriptions, characteristics, and qualities, of items to provide uniformity of meaning and reduce confusion. This chapter delineates proper nomenclature for test items.

The following topics are covered in this chapter:

- definition of a test item
- nomenclature for test items
- item formats
- purpose for test items in measurement
- criteria for good test items
- assumptions for test items
- classification of types of test items
- understanding how examinees respond to items

DEFINING A TEST ITEM

In any study, it is important to begin with a definition or scholarly description of terms. Since this book is about constructing test items, it is therefore logical to state a precise definition of the term "test item." Curiously, however, not until recently has there been proffered a uniform definition of what a test item is, although there are some antecedent, rudimentary descriptions (e.g., Gronlund, 1988; Wesman, 1971).

The lack of a uniform definition for this important term is surprising since numerous glossaries of the vocabulary of testing define many other terms associated with test items, such as item analysis, item bias, item difficulty index, item discrimination index, item characteristic curve, and more (e.g., APA/AERA/NCME, 1985; Gregory, 1996; Sax, 1989). Even an accurate description for the term test—which is characterized in many glossaries as a collection or set of test items—is not possible until the definition of test item has been established. To address this serious omission in the field, a complete and technically precise definition of a test item is offered here. It is hoped this definition will be adopted for use in the field.

The Definition of an Item

Osterlind (1990a) offered this definition of a test item:

> *A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form for answering; and, it is intended to yield a response from an examinee from which performance in some psychological construct (such as an knowledge, ability, predisposition, or trait) may be inferred. (p. 3)*

The definition is comprehensive because it includes all of the requisites for a test item regardless of whether a particular item is used for psychological assessment or educational measurement, and it is applicable to all item formats. The definition is limited to test items used in tests of achievement, aptitude, or ability. Tests of this sort are used in clinical testing, in educational and psychological testing in schools, in counseling, in employment settings, and in professional and occupational licensing and certification. Other types of tests, such as certain personality inventories, quantifiable data gathered during interviews, and even certain types of essay formats, do not contain the kind of test items covered by the definition.

Still, test items of the type covered by the definition are not limited to strictly paper-and-pencil inventories. For example, they may also be used in tests of physical performance for measuring psychomotor skills, or in measuring oral communication abilities. Constructed-response formats for test items are also covered by this definition. Additionally, the definition encompasses test items in the kinds of tests mentioned above regardless of whether a particular test is standardized or non-standardized, administered in a group setting or administered individually.

Understanding the Definition

The first aspect of the definition—"A test item . . . is a unit of measurement . . ."—concerns the function of measurement. Measurement means quantification, either objectively or subjectively derived (cf. Lord and Novick, 1968; Stevens, 1946; Torgerson, 1958; Weitzenhoffer, 1951; many others). Hence, a test item by this definition leads to data that is quantifiable in some manner. It is important to grasp the significance of this seemingly obvious point because test items are intended to yield numerical interpretations. The number associated with a particular examinee's performance is meant to provide a basis for comparison, usually either against a group of peers or against a predetermined standard or criterion.

The numerical interpretation for test items is what differentiates them from instructional activities. Instructional activities are not specifically designed to yield numerical data. The primary purpose for instructional activities is as a heuristic, or helping-to-discover, device. Although test items and instructional activities differ in intention, they often cover identical subject content or psychological processes.

This leads to the next important point in the definition of a test item: "A test item . . . [has] . . . a stimulus and a prescriptive form for answering." Epistemologically speaking, a test item may be considered as etiology because it is a stimulus that causes a response. Further, the response given by an examinee to a test item is prescribed in the sense that the item guides a particular form that the answer should take. For example, in a multiple-choice test item, the test taker is directed to select from among the alternatives offered; or, in a constructed-response format, like completion or short-answer, the examinee must supply a word or phrase as a response and cannot, say, circle one of the words in the item. Even in other open-ended test-item formats, like physical performance exercises, the examinee is guided to make a specific response. This is what is meant by saying the response for a test item is "prescribed." It would violate the definition of a test item if the test taker were not directed to make a particular, predetermined kind of response.

Finally, the definition states that an examinee's response is interpreted in terms of learning something about his or her performance in a particular psychological construct. Psychological constructs are hypothesized concepts for explaining human behaviors and include the kinds of attributes or traits that are the object of assessment by psychological and educational tests, such as reading ability or emotional development. Since a psychological construct is something that is only theoretically conceived and cannot be directly observed, it would be useful to have a way to infer at least the existence of a psychological construct and the relative degree to which it may be exhibited by a particular examinee. Test items perform this function. If a stimulus situation does not provide data that implies a psychological construct, it is not a test item, according to the definition. This issue (which begs a bit of explanation of psychological theory) will be taken up again in a later section of the chapter.

This concludes the discussion of the definition of a test item. The reader is encouraged to review this section thoroughly since much of what follows presumes an awareness and comprehension of the definition of a test item.

TEST ITEM NOMENCLATURE

No Current Uniform Terminology

Surprisingly, the lexicon of test items is not well established. There is a need to identify and standardize the stock of terms related to constructing test items. The definition for a test item given above provides a useful start, but many other terms relevant to constructing test items need to be considered. A few of the most important terms are described here. The reader should study them as a specialized, technical vocabulary list.

As one can readily appreciate, a number of terms are important for constructing test items. Table 2.1 lists terms routinely used in this book as well as throughout the item construction process. Definitions for terms requiring one are provided. They should be memorized and applied consistently to promote standardization in the field and to reduce the chance for misunderstanding.

Table 2.1 Key Terms Used in Constructing Test Items

- constructed-response
- correct response
- dichotomously scored
- distractor
- examinee
- foil
- graphic
- item format
- response alternative
- response
- selected-response
- stem
- stimulus
- test item
- test taker
- text

The Term "Test Item"

A test question or stimulus situation meeting the conditions of the definition discussed in the previous section should be referred to as a *test item.* This term is the most accurate descriptor for this particular kind of technical writing. An item is a single unit in a series or collection and is specified separately. The term *test item* is broad enough to allow for a variety of item formats and item classifying categories, yet sufficiently precise to be useful for technical discussion.

Test items should not be called "questions" since a test item can assume many formats, some of which are not interrogative. For example, most completion or short-answer item formats as well as most matching item formats are not stated as interrogatives. The term *test item,* on the other hand, includes both writings that are stated as interrogatives and those that are not.

Despite its aptness as a descriptor, the term *test item* has not been exclusively employed since the early days of testing. Alfred Binet (in Herrnstein, 1971), one of the first explorers into the world of mental attribute testing, called his tasks "stunts." Examples of Binet's stunts for a six-year-old are to distinguish between morning and afternoon, count thirteen pennies, and copy a diamond shape; and, for a ten-year-old, to arrange five blocks in order of weight and draw two designs from memory (Binet and Simon, 1917). Activities like these have also been referred to as "tasks." Even today, the National Assessment of Educational Progress (NAEP) labels the activities included in the NAEP program "exercises" (Messick, Beaton, and Lord, 1983). NAEP's mislabeling is unfortunate because the NAEP program has wide exposure in the popular media and may inadvertently promulgate idiosyncratic terminology. Regardless, the term *test item* is the most accurate descriptor of writings that meet the conditions for the definition discussed in the previous section,

The Examinee

The individual who takes the test is referred to as an *examinee.* Examinees are also called *test takers*; however, in most academic contexts, such as during test development or in scholarly research studies of tests or test use, examinee is the preferred term. The examinee takes the test by choice, direction, or necessity and is the most important participant in the testing process.

Regardless of the use intended for the test scores (e.g., assessment of achievement, diagnosis for counseling in clinical settings, licensing and certification, etc.), the welfare of the examinee should be primary in making decisions about tests. The *Standards* (AERA/APA/NCME, 1985), recognizing the paramount role of the examinee in the testing process, devotes an entire section to protecting the rights of examinees. In later sections of this book, too, we shall address various issues concerned with protecting examinees' rights.

Specifying Item Formats

As previously mentioned, a variety of *itemformats* are available to the writer of test items. (The term *item formats* is sometimes abbreviated to, simply, *itemforms.*) The format for a test item is simply its design and layout. Some of the most repeatedly seen item formats are multiple-choice, true-false, matching, sentence-completion, short-answer, and less frequently, cloze-procedure. Most of these latter formats may be loosely organized under the more general category of "constructed-response" or "performance" (cf. Chapter 8). Illustrative Items 2.1 to 2.6 are examples of each of these item formats, respectively. These item formats are the ones most often employed in many popular tests of mental attributes. Readers should become familiar with the formats so as to correctly identify a particular one when confronted with it.

Illustrative item 2.1 An example of the multiple-choice format.

There is an 80% chance of snow tonight. Which is the most reasonable interpretation of this forecast?

A. It will snow tonight.
•B. It will probably snow tonight.
C. It will probably not snow tonight.
D. 20% of the area will not receive snow.

Illustrative Item 2.2 An example of the true-false format.

Spanish sympathizers, in an underground movement, provided assistance to the American colonists during the Revolutionary War.

True False

Illustrative Item 2.3 An example of the matching format.

Match the numbers of the categories on the left with the corresponding letters of the characteristics on the right.

1. SENSATION	a. condolence
2. AFFECTIONS	b. rocks
3. SPACE	c. incombustability
4. PHYSICS	d. hearing
5. MATTER	e. interval

Illustrative Item 2.4 An example of the completion format.

The ____________________ branch of government is the only branch empowered to pass spending appropriations.

Illustrative item 2.5. An example of the short-answer format.

In what city was the Declaration of Independence signed?

____

Illustrative item 2.6 An example of the cloze-procedure.*

Bridges are built to allow a continuous flow of highway and traffic across water lying in their paths. But engineers cannot forget that river traffic, too, is essential to our economy. The role of ____1____ is important. To keep these vessels moving freely, bridges are built high enough, when possible, to let them pass underneath. Sometimes, however, channels must accommodate very tall ships. It may be uneconomical to build a tall enough bridge. The ____2____ would be too high ....

1. a) wind
   b) boats
   c) weight
   d) wires
   e) experience

2. a) levels
   b) cost
   c) standards
   d) waves
   e) deck

*From DRP Handbook (p. 2) Touchstone Applied Science Associates, 1986, New York: The College Board.

Table 2.2 lists several popular tests classified by the type of item format principally used in that test. From the information presented in the table, one can see that although the multiple-choice form is the item format that is most popularly used, others are common. In fact here are still more item formats besides those listed above, including combinations of common item formats. Some of these other item formats are quite complicated and have been invented to serve specialized purposes and are not widely used.

Table 2.2 Tests Classified by Item Format Principally Used

Test	Item Format Principally Used
Analysis of Learning Potential (Harcourt, Brace, & World)	Multiple-choice, completion, and analogy
California Achievement Tests (CTB/McGraw-Hill)	Multiple-choice
Cognitive Abilities Test (The Riverside Publishing Company)	Multiple-choice
College BASE (University of Missouri-Columbia)	Multiple-Choice and writing
Comprehensive Tests of Basic Skills (CTB/McGraw-Hill)	Multiple-choice
Cornell Critical Thinking Test Level X (Illinois Thinking Project, Univ. of Illinois)	Matching
Degrees of Reading Power (College Board, The Psychological Corp.)	Cloze-procedure
Graduate Record Examinations (Educational Testing Service)	Multiple-choice, completion, and analogy
Iowa Tests of Basic Skills (The Riverside Publishing Company)	Multiple-choice
Kuhlmann-Anderson Tests (Scholastic Testing Service)	Multiple-choice

Table 2.2 *(continued)*

Test	Item Format Principally Used
Metropolitan Achievement Tests (The Psychological Corp.)	Multiple-choice and writing
Metropolitan Readiness Test (Harcourt Brace Jovanovich)	Multiple-choice with pictures
Miller Analogies Test (Prentice Hall Press)	Multiple-choice
Minnesota School Attitude Survey (MSAS) (Science Research Associates)	True/false, Important/unimportant
National Registry of Radiation Protection Technologists (NRRPT)	Multiple-choice
Otis-Lennon Mental Ability Tests (Harcourt, Brace, & World)	Multiple-choice
Otis-Lennon School Ability Advanced Form S (The Psychological Corporation)	Multiple-choice with pictures
School and Ability Tests (Addison-Wesley Testing Service)	Multiple-choice and analogy
Secondary Level English Proficiency Test (CTB/McGraw-Hill)	Multiple-choice
SRA Achievement Series (Science Research Associates)	Multiple-choice
Survey of Basic Skills (Science Research Associates)	Multiple-choice
Tests of Achievement and Proficiency (The Riverside Publishing Company)	Multiple-choice

Table 2.2 *(continued)*

Test	Item Format Principally Used
Test of Cognitive Skills (CTB/McGraw-Hill)	Multiple-choice
Tests of Adult Basic Education (CTB/McGraw-Hill)	Multiple-choice
The Gifted Evaluation Scale (Hawthorne Educ. Service)	Likert-type rating scale
The Stanford Achievement Test (The Psychological Corp.)	Multiple-choice

Still other tests may contain several item formats in the same test instrument. An example of a test with a variety of item formats, as well as other assessment techniques, is the *Alabama English Language Proficiency Test.* The various formats for items used throughout this test are described in Table 2.3. Notice in the test that not only are traditional item formats used, but there is also a portion in which examinees listen to a recorded stimulus and respond by answering a set of multiple-choice test items.

Caution is advised, however, when combining various item formats into a single measure. The interpretability of the scores yielded by such complex assessments can be suspect and may sometimes even be spurious. Wainer and Thissen (1992), in a research report provocatively titled "Combining Multiple-Choice and Constructed Response Test Scores: Toward a Marxist Theory of Test Construction," provide a technical discussion of some of the measurement issues that may arise when multiple-choice and constructed-response formats are combined in AP (Advanced Placement) tests with high school students. And Downing (1995) discusses evidence for valid score interpretations in a medical specialty certification examination comprised of mixed item formats, including multiple-choice and multiple true-false.

Table 2.3 Example of a Test in Which a Variety of Formats is Used

Alabama English Language Proficiency Test

Content Area	**Assessment Method**
Reading	A cloze test of reading comprehension, using multiple-choice items with five choices.
Writing	An essay test, scored by the holistic method.
Language Skills	A multiple-choice test (four choices per item) of basic grammar, mechanics, and reference skills.
Listening	A listening tape of passages read aloud, testing comprehension by multiple-choice items.

Constructed-response: Writing Samples, Exercises, and Essay-Type Questions

Forms in which the examinee actually constructs a response (as opposed to selecting one alternative among several) such as writing samples, writing exercises, and essay-type questions may also be considered item formats because this type of assessment meets all of the conditions for test items described in the definition given earlier. They are a stimulus situation, they have a prescriptive form for response, and they are intended to yield scores that allow for inferences to be made about examinee performance in a psychological construct. However, since they are a specialized format and different from the other item formats discussed in much of this book, they are given a chapter all their own. There, we will explore this format in depth. Readers interested in learning more about writing samples specifically and techniques for scoring them are referred to the fine book by White (1985) which seems destined to become a classic of the genre.

Types of Item Formats

Test item formats fall into two broad types: *selected-response* or *constructed-response* (or by some authors, *supply-type*). In a selected-response test item the examinee is given the correct solution to the problem as well as alternative (and usually incorrect) solutions. The examinee is instructed to select the perceived correct answer. Multiple-choice and true-false test items, the most commonly used item formats, are selected-response test items. In these formats, the examinee is instructed to choose one response alternative from among those offered. Illustrative Items 2.1, 2.2, 2.3, and 2.6 are examples of the selected-response item type.

By contrast to selected-response test items, constructed-response test items alternative solutions (correct or incorrect) are not presented to the examinee at all; rather, the examinee must furnish (or "construct") the perceived correct response. Typically, the examinee responds to constructed-response test items by writing down a word or short sentence perceived to be the correct response. The completion or short-answer test item is an example of a constructed-response test item, but they may include more extended responses. Illustrative Items 2.4 and 2.5 present examples of uncomplicated constructed-response items. Chapter 6 is devoted to performance assessment and provides many more examples of constructed-response formatted items some of which present complex scenarios.

Terms for Parts of the Test Item

The part of the test item which asks the question or sets up the situation for response is referred to as the *item stem.* An item stem is thought of as a *stimulus,* because as etiologies, they cause a response from the examinee. A *correct response* is elicited when an examinee selects, identifies, or provides the answer to the stimulus that is scored positively.

In selected-response test-item formats, the choices provided are labeled *response alternatives.* (Less formally, response alternatives may be called *options*.) The response alternatives offer all of the possible choices that the item writer has provided (correct and incorrect) to the examinee. Response alternatives that are not considered the correct response are labeled *distractors.* The term *distractors* is used because these responses may distract, or daunt, the examinee who is uncertain about the correct response. In England, distractors are often called *foils*, but in the United States the term *distractors* is more accepted.

The multiple-choice test-item format (and to a lesser degree some other test-item formats) is often accompanied by narrative or graphic material. This narrative may be a paragraph or passage from a story (either originally written or excerpted from a longer work), a poem, an article or editorial from a newspaper or magazine, or other such textual material; or, it may be a graphic, such as a cartoon, map, chart, graph, table, or formula. The nomenclature for describing these accompaniments to test items is straightforward. If the type of material is narrative, it is referred to as *text.* If the material is anything other than text, it is called a *graphic.* Mathematical formulas, symbols, geometric shapes, and algebraic expressions are also considered graphics because in many printing operations they are not offset or typeset by the same procedures as is used with text; rather, they are treated as though they are pictures. In some computerized page-layout operations, such figures are contained in a formats called PICT to TIFF.

It is important to realize that when an item contains text or graphic materials, the text or graphic is an integral part of the test item and not a mere addendum. The care spent on preparing the text or graphic should equal the care used in constructing other parts of the item. Chapters 5, 6, and 7 discuss these features in detail.

Terms Used in Scoring Test Items

Another important term is regularly used in constructing items, although it does not refer to a specific part of a test item. The term *dichotomously scored test items* is used for identifying and classifying test items. The classification of test items as *dichotomously scored* means that an examinee's response is considered to be in only one of two possible categories, usually either *correct* or *incorrect.* The "correct" response has been predetermined by either the writer of the test item or some clearly established methodology. Obviously, a response other than the correct response is considered "incorrect." Most multiple-choice, true-false, matching, completion or short-answer, and cloze-procedure test items are dichotomously scored.

Although responses to dichotomously scored test items are usually categorized as correct and incorrect, other categories for responses can also be used. For example, sometimes an examinee is directed to respond to a test item with either *agree* or *disagree,* as in the examples given in Illustrative Items 2.7 and 2.8. These test items are also dichotomously scored although there is no correct or incorrect response alternative.

Illustrative Item 2.7 & 2.8

Read each statement below and decide whether it conforms to your personal sentiments. Indicate your opinion by circling one of the choices provided.

2.7. Children should be seen and not heard. AGREE DISAGREE

2.8. Breaking a mirror will bring seven years bad luck. AGREE DISAGREE

Test items that are scored dichotomously are sometimes called *zero-one* test items after the representational computer binary logic in which examinees' test scores are sometimes retained in computer records as a "0" to indicate an incorrect response and a "1" to signify a correct response. (In Illustrative Items 2.7 and 2.8 a "1" would indicate "Agree" and a "0" would symbolize "Disagree.")

Not all test items are scored dichotomously. For example, test items that prescribe multiple response options, many (but not all) short-answer test items, and some other test-item formats are not scored as having only two options for response. Often, these test-item formats allow for a variety of responses that could be considered appropriate to the question. Further, there are many models for scaling tests to yield numerical interpretations of the scores. Some of these scaling models are *polychotomous,* meaning that for an item more than one response alternative is considered in scoring, others are heuristic, and still more may be categorized in the psychologically based scaling methods of item response theory. For practical purposes of learning about constructing test items, however, it will be convenient to consider only items that are dichotomously scored.

Nearly all tests that are electronically scored by an optical scanning device contain only dichotomously scored test items. Some exploratory work is being done to enable optical scanners and their concomitant computers to read multiple response formats, including some kinds of essay examinations. Presently, the application of this work is quite limited, but such work does promise exciting vistas for new testing formats in the future.

Putting Together the Parts of an Item

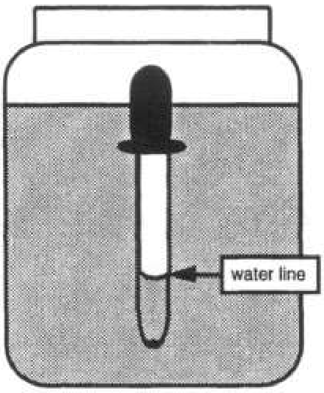
Thus far, we have discussed several constituent parts of test items, including the stem, response alternatives, text, graphics, and more. Now let us see how they appear in an item. Figure 2.1 notes the constituent parts of a multiple-choice test-item format. Study the placement of the various parts in the item. With few exceptions, the arrangement of parts of a test item will remain fixed. The order is that the directions appear first, followed by a graphic (if any), which is followed by text (if any), which is followed by the stem, and finally the response alternatives.

Now, with criteria, terminology, plenty of examples, and some understanding of items in place, we are ready to look at the big picture: How does an entire item appear? Figure 2.2 presents a multiple-choice test item dissected into desirable parts. This figure should provide the reader with a good notion of what test items are all about in a general sense—a sort of *Gestalt* perspective.

Figure 2.1 Test item nomenclature.

Use the illustration and read the passage below to answer the question.

Direction



Graphic

Charley did an experiment in which he floated an eyedropper in a jar of water. He left the jar on a shelf for one hour. Then he recorded the level of water in the eyedropper. Next, he heated the jar containing the eyedropper for 30 seconds.

Text

Which variable did Charley change in this experiment?

Stem

A. kind of eyedropper

B. level of water in the jar

C. amount of time between temperature measurements

Distractors

•D. temperature of the air and water in the jar

Correct Response

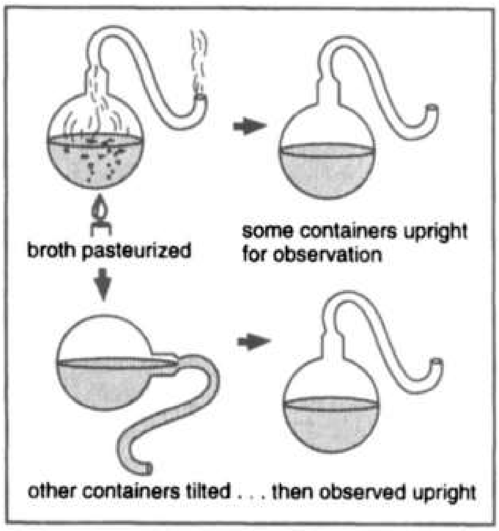
Figure 2.2 Anatomy of a test item.

Read the passage and use the illustration to answer the question.

Clear directions guide examinee.

In the nineteenth century, Louis Pasteur performed an experiment in which he bent the necks of flasks into “S” shapes, leaving their ends opened. Then he boiled broth in the flasks to force air out and kill any microbes inside. After the flasks cooled, he left some of them upright for observation. Before setting aside others to observe, he tilted them so that the broth moved up into the bent necks and then back into the flasks. After the flasks had been prepared, he watched them for signs of microbial growth.

Wording is preciseand succinct. Grammar is correct, following rules of composition.


broth pasteurized
some containers upright for observation
other containers tilted . . . then observed upright

Appropriate graphic supports item without giving undue clues.

Which hypothesis was Pasteur testing in this experiment?

Item stem asks reasoning type question (not just recall offacts).

A. Flasks with bent necks would cause microbes to grow in the broth.

B. Cooling broth in the flasks would cause microbes to grow in the broth.

C. Heating broth in the flasks and then cooling it would cause microbes to grow in the broth.

D. Contact of the broth with something in the necks of the flasks would cause microbes to grow in the broth.

Distractors are plausible; none can be rejected out-of-hand. A single, clearly correct response.

PURPOSE FOR TEST ITEMS

Psychological Theory as Background to Items

Since the primary purpose for test items is embedded in psychology, a bit of background in behavioral theory is needed before the function of test items in measurement can be explained or appreciated. Psychologists observe human behavior, and when a pattern of behavior is performed consistently over time and in different contexts by many individuals, it is labeled as a *psychological construct* (Cronbach, 1971; Cronbach & Meehl, 1955; Messick, 1975). Accordingly, a construct is a psychological attribute which characterizes individuals' behavior. Since we cannot know exactly what processes are occurring in another's brain, these psychological constructs are only hypothesized, or theoretically imagined. There may be a countless number of them for every individual, explaining the incalculable behaviors people exhibit.

Psychological constructs are such things as verbal or quantitative ability, social or emotional development, reasoning ability, spatial visualization, and the like. Endurance is a frequently used construct in athletics. Such constructs can be hypothesized as explaining broad behaviors, as in the list of constructs just cited, or they may be more narrowly specified, such as vocabulary development.

Psychological constructs are often contrasted with physical attributes such as height, weight, or the color of skin, hair, and eyes. Physical attributes are directly observable, and measuring them is often comparatively easy. Generally speaking, the instruments used to measure physical attributes yield more reliable scores than those employed in assessing psychological constructs. For instance, one does not typically worry about the reliability of a yardstick. And, for most common uses, one yardstick is as good as another.

By contrast to the relative ease of measuring physical features, assessing psychological attributes is challenging at best. There are two important reasons why this kind of measurement is difficult. First, since psychological constructs cannot be observed, they can only be assessed indirectly. The practical consequence of measuring human behavior indirectly is that the instruments used for the assessment are themselves suspect. The reliability of any particular test score of mental attributes can only be *estimated* rather than absolutely known. Further, these tests vary dramatically in quality, and as contrasted to the similarity of most yardsticks, one test cannot be haphazardly replaced by another.

The second reason for difficulty in assessing mental abilities is that a psychological construct is both subject to change and only vaguely understood. It is known, of course, that people can determine and change their behavior volitionally, thereby influencing a construct, positively or negatively. People learn to read, for instance, or juggle balls, or memorize mathematical formulas, or whatever. Thus, a construct itself can change between measurements, complicating reliable measurements. As mentioned, psychological constructs are only theoretically conceived, and comparatively little is understood about them.

As a parallel to measurements made in the physical world, imagine trying to use an elastic ruler to determine the diameter of a cloud, whose physical properties are only vaguely understood, and whose dimensions keep changing! Measuring psychological constructs is indeed difficult. Despite the formidable difficulties of measuring psychological constructs, tests of mental attributes are the main means by which one may objectively learn the psychological composition of an examinee.

Items as Measures of Constructs

With this brief background in psychological theory, one is ready to learn the primary purpose for test items, but first a note about terminology which should make the subsequent discussion a bit easier to follow. In this discussion the terms mental attribute and psychological construct will be used interchangeably. The term ability will be used in descriptions of the degree to which one possesses or exhibits an attribute or construct.

Simply stated, the dominant purpose for test items in the kinds of tests discussed here is to function as a focused stimulus which elicits a response from a particular of a psychological construct. Through test items a psychological construct is operationally set forth as a behavior which an examinee is instructed to exhibit. Such behavior might be to spell a particular word correctly, perform some mathematical operation, or rotate a cube in a specified manner. Of course, the behavior could be any number of things depending upon which construct is being assessed. By responding to the stimulus of a test item, an examinee exhibits behavior from which one may infer the existence of a psychological construct.

So far, this discussion has been relatively straightforward. The situation gains complexity, however, when the dimension of assessing the *degree* to which a psychological construct exists in an examinee is added. E. L. Thorndike (1904), an early proponent of measuring mental attributes, stated that whatever exists at all exists in some amount. Although the existence of psychological constructs is only inferred, it is logical to presume that they must also be present in some amount. Further, since psychological constructs are mental attributes, individuals will possess them in varying amounts, or degrees. Again, test items are the means by which the relative degree of a psychological construct is assessed.

The reader will recall that earlier it was stated that test items are, by definition, a unit of measurement. Therefore, by observing examinee responses to a particular set of test items that exhibit a positive correlational relationship to a specific psychological construct, it is possible to estimate how much of the construct or mental attribute an examinee may possess. It is presumed that for dichotomously scored test items, examinees who respond correctly are exhibiting a greater degree of the particular construct than examinees who do not respond correctly.

Constructs Exist for All Persons

A subtle but important point should be noted in the general case for test items just stated. The inference made is that examinees who respond correctly to items are said to possess *more* of an attribute than those who do not respond correctly. But it is not a correct interpretation to suggest that those who do not respond correctly to items do not possess the construct at all.

A construct, once hypothesized, exists for all persons, regardless of whether they can exhibit behaviors associated with it. Some persons may possess much ability in a construct, while others may have more limited ability. For example, an illiterate person of normal intelligence still possesses the construct "reading ability," since it has been hypothesized to exist in all persons of normal intelligence. However, since the construct has not been developed in the illiterate individual, this person would not be able to perform behaviors (i.e., respond correctly to test items) from which the existence of the construct could be inferred. Presumably, with tutoring assistance and practice, the construct could be developed in the person, after which he or she would likely perform the behaviors requested in test items. By contrast, psychologists have not hypothesized the reading-ability construct to exist in chimpanzees, and no amount of tutoring assistance or practice could cause a chimpanzee to exhibit the behavior of responding correctly (above random marking) to a set of test items designed to assess reading comprehension.

Although the general case for classical test theory was stated above, in many particular instances, an examinee's having additional items correct does not always indicate greater ability in the construct. This is because the complexities of scaling tests, as well as error in measurement and other factors, can make the interpretation of test scores quite complex. For example, tests that are scaled by models of item response theory as well as other polychotomous scaling models use a different set of assumptions than does classical test theory and they may not follow the theory described above. This is why it is important to realize that in interpreting test scores there is not a strictly linear relationship between an examinee responding correctly to test items and the degree to which that individual possesses ability in the construct.

This point may be illustrated with a simple example. Suppose two test items representing a construct were administered to two examinees, one of whom responded correctly to one item and the other responded correctly to both items. It could not be inferred from comparing the performance of the two examinees that the examinee who got both items correct possesses twice the ability in the construct than the examinee who got only one item correct. In fact, all that is known from this simple example is that two examinees exhibited behaviors in the same psychological construct. It is likely that the high-scoring examinee possesses *more* of the ability, but improbable that this person possesses *twice* the ability of the lower-scoring examinee. With only the information given, nothing is known about the *relative degree* to which either examinee may possess the psychological characteristic of interest. A great deal of additional information about the specific measurement instrument used, how it is scored and its scores scaled, the procedures used for selecting the examinee population, and other factors is needed before one may correctly interpret the degree to which examinees possess a particular construct.

In sum, then, test items provide a way for examinees to perform behaviors from which one may infer the degree to which a psychological construct exists in those examinees. This is the function of test items in measurement, and it justifies the earlier claim that test items are the backbone of measuring mental attributes.

CRITERIA FOR GOOD TEST ITEMS

Difficulty of Establishing Criteria

Criteria for good test items are difficult to state in absolute terms. Simple lists of declaratively "good" and "bad" features are inadequate to the task. And, as is emphasized throughout this book, an idiosyncratic list of "dos" and "don'ts" is woefully poor for what we are about here. Constructing good test items involves more than mechanically checking off considerations one by one, and such a simplistic approach will not reflect the complexity of the task. It is difficult to establish strict criteria for good test items for a number of reasons.

One reason for difficulty is that constructing test items demands technical skills. The technical skills required for constructing test items are quite complex and demand sophisticated levels of thinking. Major portions of this book provide a synthesis of the technical skills needed to construct test items.

In addition to technical competence, the skilled item constructor must also possess a penetrating understanding of how examinees respond to test items, including an awareness of how a particular test item may be perceived by different examinees. Internalizing these aspects of examinee performance will assist one in gaining a "sixth sense" about constructing test items. This deeper understanding can foster original and imaginative thinking about test items and will help one become a better writer of test items. Popham (1984) describes writing test items as "art, art, and more art" (p. 40). While this aphorism may be exaggerated, it does point to the element of creativity in constructing test items.

As a simple exercise, after reading this chapter, try to write one or two test items yourself about any subject. You may well realize immediately that in addition to needing a fundamental core of knowledge about test items, you "feel" the need for the creative component. Technical skill, coupled with a creative sense, is necessary if one is to construct good test items.

Another factor that makes it difficult to specify criteria for constructing good test items relates to the specific circumstances in which a test arises or the purposes for the test. Those circumstances or purposes dictate that the test-item writer follow certain guidelines. For example, suppose a test is being constructed as a licensing examination for paramedics who have been instructed in a standardized training curriculum. The items in such a test must assess specific criteria for service in the field by health-care personnel without regard to how a particular individual may have achieved such skills: whether by prior experience, through a required training program, or through self-study. It is the end result that counts in this context.

This is a completely different context from, say, a test that is intended to be used for assessing basic cognitive processes in a group of partially sighted children who have received varying amounts of attention to their individual learning needs. The differing contexts will require different considerations in writing the items.

Accepted Criteria for Good Items

Despite the difficulties mentioned above of specifying criteria for constructing good test items, certain criteria for good test items have been generally accepted. These criteria, first articulated by Osterlind (1990b) are standards that should be followed during construction of test items.

The first criterion for constructing good test items, and the most important, is that there must be a high degree of congruence between a particular item and the key objective of the total test. Simply stated, the primary question is, "How well does a particular test item match its intended objective?" (In this discussion, the term *objective* is being used as a synonym for a psychological construct.)

This congruence criterion is the item writer's primary consideration because it is at the heart of validity, the most important consideration in test construction. According to the *Standards* (AERA/APA/NCME, 1985), validity refers to "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p. 9). Validity is pervasive throughout the process of constructing test items and is discussed fully in the following chapter. For now, however, keep in mind that congruence between a particular test item and a specific objective (or psychological construct) is the most important criterion for constructing good test items and is related to valid interpretations of a test's scores.

A second criterion for constructing good test items is closely allied with the first. This criterion is that the key objectives must be clearly defined. If a test item is to meet the congruence criterion, it must be matched to a defined entity. To state loosely that an item measures a very broad concept, like critical thinking, without further defining what is meant by critical thinking, diminishes its potential for yielding validity evidence. The less ambiguously a domain of content or psychological construct is described, the better the chance that the congruence criterion can be met.

The notion of clearly defining the criterion or behavioral domain to be assessed may sound obvious, but it is one of the most difficult of the criteria to satisfy. Perhaps for this reason it is often carelessly ignored or only superficially addressed. Regretfully, many tests, even some published by professional test publishing companies and other professional test developers, do not adhere to this criterion for good test items.

Hambleton and Eignor (1978) proposed a set of guidelines for evaluating criterion-referenced tests. One of the guidelines asks, "Is each objective clearly written so that it is possible to identify an 'item pool'?" The authors report that this guideline was not adequately met in any of eleven popular, commercially produced criterion-referenced tests. Obviously, this is a serious deficit in test construction generally, but one that must be addressed during item construction. How one deals with this criterion in actual test development will be explored in Chapter 3, which discusses thoroughly the issues involved in considering this criterion and offers several strategies for addressing it.

A third criterion of constructing good test items is that each item's contribution to measurement error in a test's scores should be minimized to the extent possible. This means that there should be a systematic consideration of the degree to which test items may contribute to errors of measurement.

These errors may be of two types: random error (which could be caused by a variety of factors, some of which cannot even be identified with precision) and bias (the systematic distortion of measurement by over- or under-estimating a population parameter). Although the exact source of bias in test items may not be isolated, the fact that it occurs can be discovered. Following the discovery of bias in test items, the sources of bias can usually be reduced or even eliminated by repairing particular items or discarding them from further consideration. Errors of measurement, whether random or systematic, are issues of reliability, itself a special condition of validity. A thorough discussion of random errors of measurement and systematic bias, as well as techniques to detect them, is presented in Chapter 8.

A fourth criterion for good test items is that the format be suitable to the goals of the test. Straightforward, uncomplicated goals generally require simpler item formats than those necessary for assessing complex goals. Further, elaborate item formats usually consume more time during testing than simple item formats. For example, some complex item formats are inappropriate for speeded tests that are intended to cover a broad domain of knowledge under strictly timed conditions. The individual situation will provide the item writer with a context from which a careful consideration of the appropriateness of an item format may be made.

A fifth criterion for constructing good test items is that each item meet specific technical assumptions. These assumptions are psychometric concerns of measurement theory, as, for example, unidimensionality of items or local independence of items. These assumptions are delineated and explained later in this chapter.

A sixth criterion for constructing good test items is that they be well written, following uniform style or editorial standards. These standards cover grammar, diction, usage, spelling, punctuation, and syntax. While there is little disagreement among test specialists about the need for good writing in test items, editorial rules for writing test items are not well established despite some weak claims to the contrary (e.g., Cunningham, 1986). Chapters 5, 6, and 7 articulate clear editorial guidelines and an accompanying rationale for many common item formats.

A seventh and final criterion for constructing good test items is that they satisfy legal and ethical questions. Sometimes, test developers may be tempted to use another's set of well-constructed test items to save time or effort. Often, little or no consideration is given to matching the purposes of their own test. Regretfully, this form of plagiarism happens too often. Not only is this unethical, it can be illegal. When test items are copyrighted—and commercial test publishers and other test specialists frequently take pains and expense to ensure that they are—using the test items without permission infringes on federal copyright statutes.

This does not mean that all test items on every new test must be originally constructed. Frequently, the original author or copyright holder of particular test items will grant use to someone else, especially when the items are intended to be used for research purposes. Of course, test items may also be exchanged, sold, or loaned. Such use of another's work is ethical and legal if done with permission and a legal contract when appropriate.

These seven criteria for constructing good test items are summarized in Table 2.4. The careful writer of test items will become informed about these criteria and make a diligent effort to ensure that they are adhered to for each test item written. In doing so, the writer will increase the probability of writing test items that have technical merit.

Table 2.4 Criteria for Constructing Test Items

- must be congruent with key objective (or psychological construct)
- must have clearly defined key objective (or psychological construct)
- writer shoud consider the degree to which test items may contribute to errors in measurement
- the test items format should be appropriate to the goals of the test
- must meet technical assumptions for test items
- should follow prescribed editorial standards and style guidelines
- writer should review ethical and legal concerns

ASSUMPTIONS FOR TEST ITEMS

Assumptions as Theoretical Background to Constructing Items

Descriptions of modern test theories, especially those which emphasize the psychologically based nature of constructs and particularly item response theory, often emphasize relevant assumptions of mathematical models (see Birnbaum, 1968; Crocker & Algina, 1986; Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968; Thorndike, 1982; Weiss & Yoes, 1991; Wright & Stone, 1979; many others). But, items used in tests constructed by traditional, or classical, test theory should also be consistent with assumptions of mathematical models. The assumptions described below apply to all test items, regardless of whether they are included in tests developed according to classical or modern test theories. A basic awareness of these assumptions is necessary if one is to fully understand test-item construction. Furthermore, heeding them while writing items is necessary to producing good items. The three fundamental assumptions for test items are: unidimensionality, local independence, and item characteristic curves. Each of these assumptions is explained in this section.

Because this section describes items in the context of measurement theory, it may appear more difficult than some other portions of this book.

Regardless, the reader is encouraged to study this section slowly and carefully. Each issue is explained thoroughly, although, as theory, some of the points made are conceptually difficult. In fact, this may be the hardest section of the entire book to read and understand. But, an understanding of the assumptions for test items, as well as an intuitive grasp of their significance, is important for the skilled item writer.

The Assumption of Unidimensionality for Items

The first assumption underlying a mathematical model for test items is that the items are *unidimensional.* Unidimensionality of a test item means that an examinee's response to a test item can be (by inference) attributed to a single trait or ability. In other words, a test item is designed to measure one, and not more than one, psychological construct. For example, a test item that is designed to assess the trait quantitative ability measures only that trait and does not also assess other traits or abilities, such as verbal ability. Additionally, in theory, if it were possible to identify all of the possible test items for a particular construct (this would undoubtedly be an infinite number), they would define fully every aspect of the construct or latent ability.

In practice, the assumption of Unidimensionality can never be fully met because there are simply too many unknown or uncontrollable factors which affect an examinee's response, making it impossible to state with absolute certainty that a particular response was because of a specific psychological construct. Such unknown or uncontrolled factors might include the degree of motivation for the examinee, practice in responding to multiple-choice test items, familiarity with marking answers on an answer sheet, test anxiety, weariness, and many more.

Despite the difficulties in meeting fully the assumption of Unidimensionality, it is applicable to test-item construction for two important reasons. First, without the assumption, the interpretation of test items would be profoundly complex. If an item is thought to assess two abilities, there is no reliable method to infer from an examinee's response the degree to which either of the two abilities contributed to a correct response. Was the correct response due completely to the examinee's ability in just one of the two traits? And, if so, which one would it be? Or, did the examinee correctly respond to the item by drawing upon abilities in both areas? If so, to what degree did each ability contribute? By current methods of scaling, it is hopelessly complicated to attempt reliable interpretations for test items that are other than unidimensional.

Theoretical work has explored the possibility of interpreting test items in multidimensional tests; however, this work is preliminary, limited in scope, and of no practical application at this time (see Muliak, 1972; Samejima, 1974). Nevertheless, the future for work in this area seems bright. Reckase (1979; 1985; 1986) and Reckase, Ackerman, and Carlson (1988) have investigated multidimensional scaling models with item response theory and their work appears to offer enormous potential for new test scaling models and untapped test score interpretations. And, Hambleton, Swaminathan, & Rogers (1989) suggest that this approach may have consequences for the development and scaling of newer models of assessment, such as "authentic" models.

A second reason for the importance of the unidimensionality assumption is that it is widely accepted by test constructors as a conceptual notion. Osterlind (1983) noted:

> *The practicalities of score interpretation make the assumption of unidimensionality almost universally accepted by test constructors. Items from a unidimensional test may not correlate highly with each other, but only a single ability accounts for an examinee correctly responding to an item or set of test items. (p. 57)*

Two further considerations about the assumption of unidimensionality are 1) understanding that it is contextually related and not an absolute within a particular test item, and 2) understanding there are some, albeit rare, instances where it does not apply, such as in some timed tests. Regarding the first consideration, while it is convenient to think that in most circumstances all well-constructed test items are unidimensional, a given test item does not possess the characteristic of unidimensionality once and forevermore. A single test item resides in the context of a set of test items, which are governed by the purposes for the whole test. A test item may be unidimensional for one test but not for another. The veracity of the unidimensionality assumption will depend upon the purposes of the test as well as the particular set of test items used.

This concern becomes particularly acute when tests are translated from one language to another or are used with examinees whose cultural background is different from the cultural background of the group for whom it was originally written. When test items are translated literally into another language, new dimensions of meaning arise which can distort the original. The assumption of unidimensionality of any particular test item may be violated by literal translation of test items. Further, the denotations, connotations, and various nuances of language can vary markedly between different cultures even when they share a common spoken and written language.

This phenomenon may make a very good test item in one setting inappropriate in a new situation. For example, in New Guinea, English is the language of government, education, and business, just as it is, of course, in the United States. In some primitive regions of New Guinea, however, there is a different concept of time from that held by most Americans. Therefore, to ask a Papuan child to order the months of the year would be nonsensical, although it may be a very good exercise for an American youngster.

One further feature of the assumption of unidimensionality needs mention. Embedded in the assumption is the notion that test items are not administered under speeded conditions (Hambleton & Swaminathan, 1985). Speeded conditions are employed in tests that have instructions for administration requiring examinees to complete as many test items as possible under rigorous time limits. The speed at which examinees respond to items is itself a variable for measurement. Under speeded conditions examinees are not expected to complete all of the test items, especially those test items that are more difficult or are located at the end of the test; hence, not all of the test items adhere to the assumption of unidimensionality.

The importance of the assumption of unidimensionality should be apparent from this discussion. Traub (1983) has investigated the veracity of the assumption of unidimensionality of test items relative to the training examinees receive; and Hattie (1981) has compiled an extensive review of the literature on definitions of unidimensionality. Readers are referred to these sources for more detailed discussion. Still, it is not the only important assumption in understanding characteristics of test items. There are at least two more, which will be considered below.

The Assumption of Local Independence for Items

The second assumption for test items is *local independence.* This assumption is distinct from the unidimensionality assumption, although it sometimes takes a bit of study to realize the distinction. McDonald (1980a, 1980b, 1982) has investigated the equivalence of the assumptions of unidimensionality and local independence for test items.

The discussion which follows is organized into two distinct parts: a theoretical description of local independence and the practical consequences of the assumption.

Unidimensionality, it was pointed out, exists in the sense that a single psychological construct explains for each homogeneous subpopulation the assessment garnered by the test item, or set of test items. This assessment is valid only for the subpopulation who may be located at a single point along a continuum of ability, from low to high. The continuum is labeled "low" to "high" because the behavior or trait is extant in all persons, although in varying degrees. The point along the ability continuum at which the veracity, or truthfulness, of the assumption may be checked is called "local independence" since what has been said about the examinees at this point on the continuum is unaffected by other subpopulations at any other point on the continuum.

Hambleton & Swaminathan (1985) provide a mathematical definition of this assumption for tests guided by item response theory: ". . . the assumption of local independence applies when the probability of the response pattern for each examinee is equal to the product of the probability associated with the examinee response to each item" (p. 23). Osterlind (1983) demonstrated this point mathematically by considering the probability of occurrence of a given five-item pattern of responses on a test. A statistical test to check the independence of item responses for examinees is provided by Lord (1952).

The practical consequence of the assumption of local independence is more straightforward than this theoretical description. In practice, local independence means that an examinee's response on any particular test item is unaffected and statistically independent from a response to any other test item. In other words, local independence presumes that an examinee approaches each test item as a fresh, new problem without hints or added knowledge garnered from responding to any other test item.

We can see the effect of violating this assumption by examining two test items in sequence, Illustrative Items 2.9 and 2.10. The point to notice when considering the two items is that the information provided in 2.9 provides clues which can be used to answer the next item, 2.10. By correctly recognizing that one characteristic of a herbivore is worn, flat teeth in the back of the mouth (cf. 2.9), the astute examinee could immediately use this knowledge to study the graphic for 2.10 and match it to response alternative C. This makes an examinee's response to 2.9 dependent upon his or her response to the preceding item. The response to 2.10 is not similarly advantaged. The items are linked in an undesirable (and unintended) way. With the local independence assumption violated, a proper interpretation is not possible.

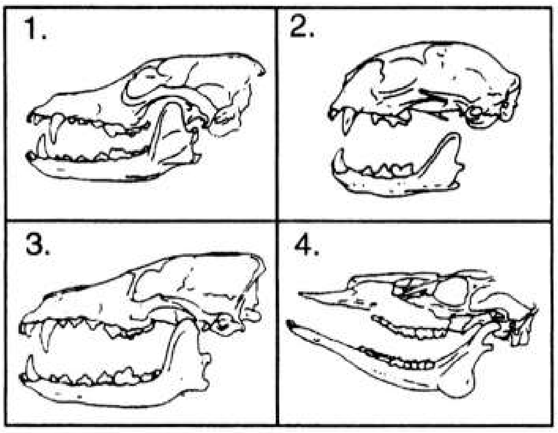
illustrative Item 2.9

Which physical characteristics is a herbivore most likely to have?

A. long, grasping tail
• B. worn, flat teeth in the back of the mouth
C. short legs with long claws
D. sharp, pointed teeth in front of the mouth

Illustrative Item 2.10

Use the sketches below to answer the question that follows.



Skull 4 probably came from which type of animal?

A. parasite
B. carnivore
•C. herbivore
D. saprophyte

One should not confuse local independence with the idea that test items cannot share a text or graphic. Having several items refer to a common text or graphic is unrelated to local independence for items. In fact, sharing a textual passage or a graphic among several items is a useful economy in item construction. It can reduce the amount of reading required and allow for more items in a test.

The Assumption of Item Characteristic Curves

A third assumption important for constructing test items concerns the item characteristic curve (ICC). The ICC is a feature of methodologies for scaling people according to their responses to the stimulus of test items. It is particularly useful for analyzing test items. In a later chapter which discusses analyzing items for quality, the practical applications of ICCs are featured prominently. Here, however, they are described in more theoretical terms, as an assumption for constructing test items. As with the two previous assumptions, the use of ICCs applies to all test items regardless of whether a particular test item is included on a test that follows classical or modern theories for developing and scaling tests.

In the language of psychometricians, ICCs represent the regression of item scores on an ability variable. Nunnally (1978) had a simpler description of ICCs that may help explain them here: "Nearly all models for scaling people can be depicted by different types of curves relating an attribute to the probability of responding in one way to items versus another" (p. 68). These descriptions for ICCs may become clearer to the reader when they are graphed and the graph studied directly. Four different types of ICCs are displayed in Figures 2.3 to 2.6.

Notice in Figures 2.3, 2.4, 2.5 and 2.6 that the attribute described in the ICC may be considered equivalent to a independent variable in that it is the thing being measured. It may be an observable behavior, a learned or applied skill, or an inferred trait. The ICC records this "independent variable" along an ability continuum, from low to high, on the abscissa, or X axis. This is the same ability continuum described earlier for the local independence assumption, but here it is graphically portrayed rather than only theoretically conceived.

Figure 2.3 An ascending linear trace line for a test item.

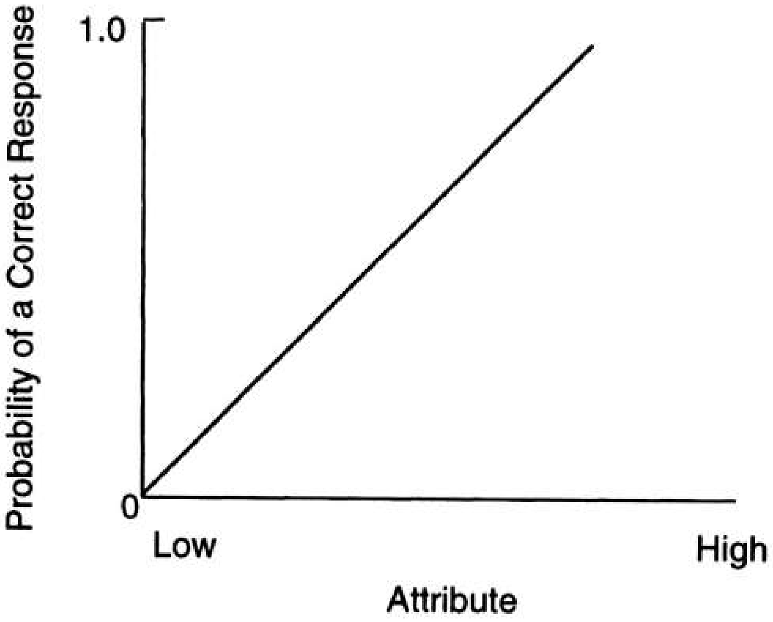


Figure 2.4 A monotonic trace line for a test item.

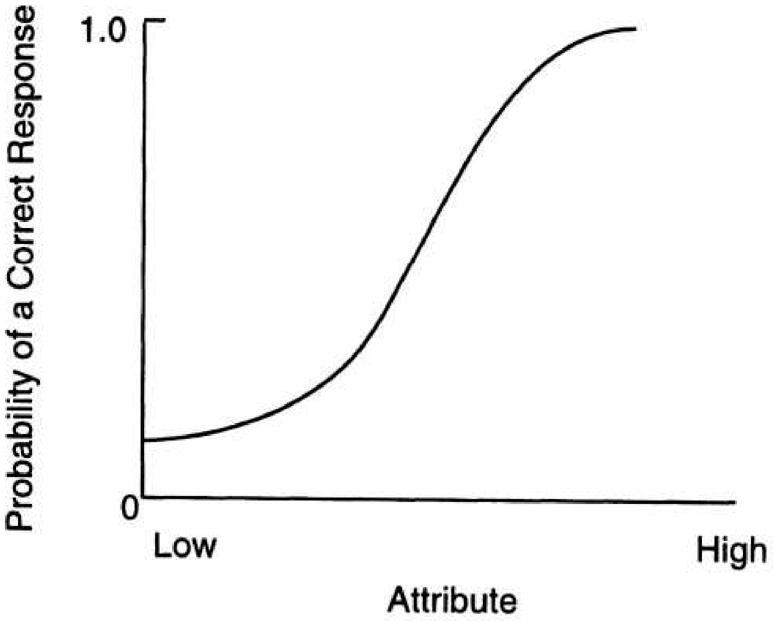


Figure 2.5 A nonmonotonic trace line for a test item.

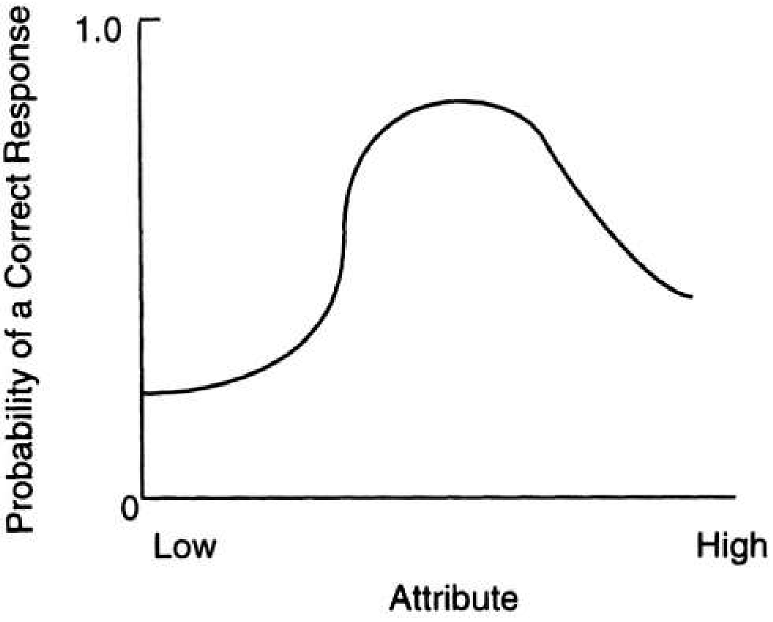
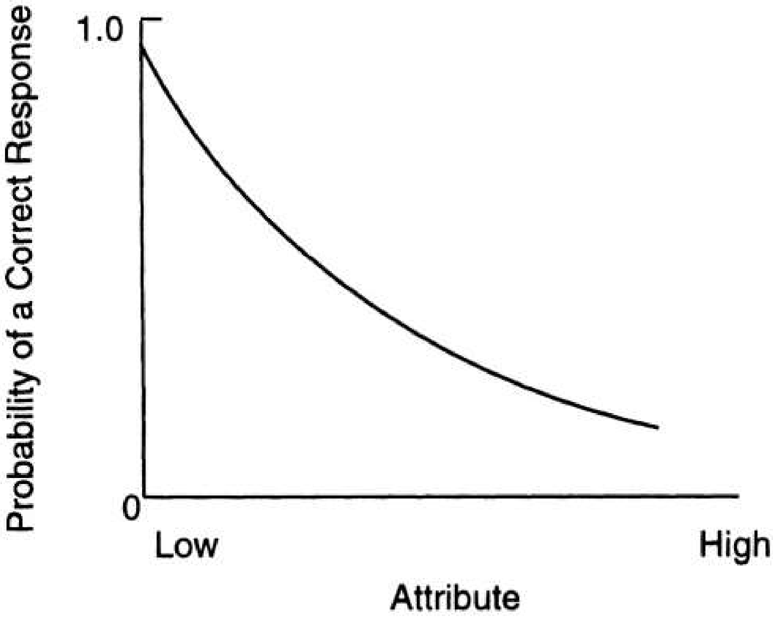


Figure 2.6 An descending monotonic trace line for a test



Next, observe in the figures that the ordinate, or Y axis, of ICCs is a measure of probability of a correct response to an item, ranging from 0, no probability, to +1, perfect (or 100%) probability. Persons possessing or exhibiting a low degree of the attribute will tend to have a small probability of responding correctly to a test item that is a valid measure of the attribute. Conversely, someone who possesses or exhibits a high degree of the attribute will tend to have a high probability of responding correctly to the same test item. Figures 2.3 through 2.6 display this relationship between ability in a construct and probability of a correct response to an item.

A correct response to a test item is expressed as a "probability" because test items are fallible, that is, they are unreliable. As is emphasized in this book as well as throughout all measurement theory, test items do not measure psychological constructs with unfailing accuracy; rather, test items permit inferences to be made about psychological constructs with a specified degree of confidence. If a test item were a perfectly reliable measure of an attribute, persons at any given ability level would have either a zero chance or a 100% chance of responding correctly to the test item. The ICCs in the figures reflect this probability.

Figure 2.3 is described as ascending, meaning that this particular ICC always increases, and linear, noting that it is a straight line because one unit of increase in the attribute means a corresponding one unit increase in the probability of responding correctly. However, in practice the relationship between ability and probability of a correct response to an item is more complex. Figures 2.4 to 2.6 each present different aspects of this complex relationship.

Figure 2.4 is described as monotonic because the item trace line does not begin at zero probability and, on the upper end, approaches, but never reaches, one, or perfect probability. This means that low-ability examinees still have some (albeit very low) probability of a correct response and very-high-ability examinees never achieve a perfect chance of a correct response. Figure 2.4 may be contrasted with Figure 2.5, which displays a nonmonotonic trace line.

Figure 2.6 displays an ICC for a poor item because low-ability examinees have a greater probability of a correct response to the item than do highly able examinees. Such a circumstance can occur when a badly worded item is taken at face value by less-able examinees but found confusing by more-able persons. This phenomenon will be examined in greater detail in Chapter 4.

As can be seen from this brief discussion, ICCs are extremely important in constructing test items. We shall return to them again at several points throughout this book. For now, however, it is important to appreciate the role they may serve in mathematically describing characteristics for assumptions of test items.

The Importance of the Assumptions

This section—certainly the most theoretical discussion of the entire book—examined three important assumptions for test items. These are the assumptions of unidimensionality, local independence, and the item characteristic curve. While understanding these concepts may require study and review, such in-depth consideration will be worth the effort. With a thorough understanding of the assumptions for test items, one realizes why particular practical steps in item construction are needed to produce good items. Additionally, the theoretical underpinnings provide a strong rationale for a consistent and uniform theory of test item construction.

CLASSIFICATION OF ITEMS

Test items can be organized in a variety of ways. One classification scheme that is especially useful organizes them according to the function they may serve in a final test instrument. By one scheme there are four functions for items, or four item types: 1) mastery-type test items, 2) power-type test items, 3) speed-type test items, and 4) independent variable-type test items. Each of the item types has a special purpose which will be described momentarily.

Some readers knowledgeable about tests may recognize that similar terminology is also used to classify types of tests. However, it should be realized that the terminology used for classifying test items does not imply that each particular item type must or will appear in a test that is similarly termed. In other words, tests that are classified as mastery, power, or speeded tests may or may not contain mastery-type, power-type, or speed-type test items, respectively. For example, a power test may contain any one type or all four types of test items. The specifications for any given test can require that one, two, three, or all four types of test items be included. The distinction between types of tests and test-item types will become clearer with a description of the test-item types.

Mastery-type test items are intended to measure essential minimums that all examinees must know. These items are typically low-level test items requiring simple memorization of facts or computation. This type of test item is commonly used in licensing and certification tests. Several airlines pilot tests contain virtually all mastery-type test items, wherein examinees are required to respond correctly to every test item. Examples of typical mastery-type test items from these tests may be to ask examinees to supply or select the correct radio frequency for control tower operations at the Los Angeles International Airport, or to determine the length of runway required for sufficient speed to achieve the necessary lift for takeoff given an airplane's thrust, total weight, and other pertinent information.

Power-type test items are designed to measure what typical or most examinees may be expected to know. These test items may range in difficulty from very easy to very hard, depending upon the subject matter and the context for the test. An example of power-type test items may be seen in a spelling test in which vocabulary words are selected for inclusion from an appropriate word list. Some words may be easy and others difficult for a given examinee or group of examinees, but most of the examinees should recognize or supply the correct spelling for most of the words. A spelling test for average-achieving fifth-graders may include words from a word list appropriate for, say, fourth- to sixth-graders. Achievement tests typically contain many power-type test items.

Speed-type test items should tap the more difficult concepts and skills that typically only the most able examinees may know. Speed-type test items should not be confused with speeded tests, in which the administration of a set of items under strictly observed time limits is itself a variable for measurement. In this context, speed relates the act of responding to test items to a theory of intelligence which postulates that intellectual ability is positively correlated with the speed with which people mentally process information (Jensen, 1980; 1982a; 1982b). This proposition, while still a theory, has strong supportive evidence and underlies most tests of mental abilities.

Independent variable-type test items have a special purpose of their own. They are designed to measure achievement in subject matter areas in which the content is evolving. For example, in the medical field new drugs typically emerge and gradually replace older ones. Often there is divided opinion and debate about which of two or more drugs (singly or in combination) may be indicated (or contraindicated) for a particular set of symptoms. Consensus among physicians, chemists, or others is slow to emerge because side-effects and consequences of drugs sometimes take years to develop. Examinees' familiarity with new drugs may be tested with independent variable-type test items.

In some circumstances independent variable-type test items are incorporated into a test instrument but not included in an examinee's score. Also, on occasions in which parallel test forms are being developed concurrent with a test's administration, it may be convenient or necessary to include test items for trial purposes as independent variable-type test items. This practice is relatively common, particularly in large-scale assessment programs in which there are parallel forms required from one test administration to the next. Again, such test items usually do not contribute to an examinee's total test score.

As may be guessed, the name independent variable-type test item is derived from research methodology wherein an independent variable is considered to be the presumed cause and the dependent variable the presumed effect. While independent variable-type test items are not in themselves the independent variable of a research methodology, conceptually they serve a loosely parallel purpose, hence, the same term is used.

Understanding How Examinees Respond to Items

Although it is superficially appealing to imagine that examinees respond to test items on the basis of either complete knowledge of the information requested or lack of complete knowledge, in fact the situation is more complicated. It has long been recognized by test developers and psychometricians that examinees respond successfully or incorrectly to items for many reasons, including complete information, partial information (of the stem or one or more of the response alternatives), misinformation, blind guessing, as well as a variety of other reasons.

Hughes and Trimble (1965) identified six combinations of information which might influence an examinee's response to an item. These are

- positive correct information which leads to a successful response,
- partial information which leads to a successful response,
- complete lack of information wherein an examinee's response is a blind guess,
- partial information which leads to an incorrect response,

- positive incorrect information which leads to an incorrect response, and
- an examinee's belief that the item is trivial, inane, or factually incorrect, and although the "correct response" is surmised, the examinee opts not to select or supply it.

This information is important for item writers to consider when preparing items for two reasons. First, by being aware of how examinees respond to test items, the writer can more suitably tailor an item to achieve a desired end. Using the same thinking strategies adopted by most examinees, the writer can read a freshly-prepared item and imagine each of these six combinations of information brought to the item by an examinee for insights into the wording proposed for the item. Second, in a general sense, the more knowledgeable the writer is about examinees as an audience for this particular kind of technical writing, the better chance he or she has of reaching it. As Wainer, Wadkins, and Rogers (1983) point out, producing high-quality test items "involves the consideration of every possible interpretation of the item" (p. 3). Good item writers must certainly put forth an extra measure of effort to ensure that the items they produce are of high quality.

CONCLUSION

In this chapter a number of concepts important to constructing test items have been introduced, including the definition of a test item and associated terminology, an explanation of the purpose for test items in measurement, a description of item formats and some examples, an exposition of criteria for good test items, a discussion of relevant assumptions, and a listing of types of test items. Finally, a brief discussion was offered of the combinations of information an examinee may bring to an item. The reader should master these concepts because the following chapters build upon the information that has been presented here. In the next chapter we will explore the idea of validity in test items by focusing on issues and strategies for determining the content of test items.

Chapter 3

Determining the Content for Items: Validity

INTRODUCTION

Determining the content for test items can be a perplexing experience. The item writer may search through textbooks or encyclopedias, curriculum guides, or other source material for grist to put into his or her item-writing mill, only to find that the result is an item whose content is trivial, pedantic, or arcane. Further, the item writer must not only find subject matter that is above the inconsequential, but in order to write items that tap complex cognitive processes, he or she must also judge the level of mental processing required of examinees to respond to a particular item. It is important to appreciate this point because determining appropriate content for items requires a complex view of items as measures not just of subject content but of psychological processes as well.

Without proper guidance, the intricate considerations of subject content and psychological processes for items are likely to result in a haphazard guess at what content is appropriate and which cognitive skills a particular item may tap. Untutored item writers may discover that what superficially appears to be a simple process of finding content often turns into a frustrating search for something usable.

Selecting appropriate content to use in constructing a test item requires more than a review of curriculum sources and certainly more than blind hope that an item taps a specified level of cognitive processing. It requires an understanding of what an item is, as well as a working familiarity with criteria for good items (these were discussed in the preceding chapter). Further, it necessitates a consideration of how the item may contribute to evidence of validity for an entire test. In fact, determining appropriate content for items is, in a very real sense, a consideration of validity. Because of the importance of validity to tests generally, this chapter focuses upon issues of validity as they relate to determining the content for individual test items.

Although validity refers to supportable interpretations of scores yielded by a whole test and not to single items, inferential interpretations for tests are possible only because the stuff of validity is imbued in the individual items. Single test items contain the "genetic material" which can bring to life supportable interpretations of measurement, and hence, test validity. Ebel (1983) even goes so far as to state that careful item construction is *de facto* evidence for validity. And Haladyna (1994), articulates three aspects of item construction as they relate to construct validation, including, "(a) the sampling of content, (b) the measurement of higher level thinking, and (c) the differentiating of test scores when the response format involves recognition versus production" (p. 27).

Before one can appreciate the relationship between constructing an item and validity, one should have an unambiguous understanding of the concept of validity. To this end, this chapter begins with an explanation of basic concepts of validity. A discussion of the relationship between constructing test items and validity follows. This information provides the necessary background material from which careful item writers can make informed choices about the content of test items.

When determining whether the content for a particular test item may contribute to eventual evidence for validity, the skilled item constructor must consider several factors. First, the item writer must have a clear conception of the test's purpose, and must work from an exactly defined domain of content or psychological construct. Second, in order to determine appropriate content for items, the item writer frequently needs to have at hand carefully crafted test specifications and possibly even item specifications.

Finally, skilled item writers also need to be mindful of systematic methods for reviewing the congruence between an item and the skills or cognitive processes that are intended to be assessed. By attending to these considerations, the knowledgeable item writer can assure that the content of particular items both matches the curricular goals of the test and assesses the intended psychological processes. This chapter describes these considerations and explains strategies for dealing with them during the item-construction process.

The following topics are covered in this chapter:

- basic concepts of validity
- relationship between constructing test items and validity
- conditions for items to contribute to validity
- initial considerations when selecting content for items
- achieving clarity in a test's content
- developing test content specifications
- melding cognitive processing levels with item content
- item specifications
- consistency of an item with its specification

BASIC CONCEPTS OF VALIDITY

Understanding Validity

The concept of validity is the paramount concern in test item construction, and understanding it is an important prerequisite to writing good items. One can begin to understand validity by careful scrutiny of a definition. A commonly cited definition of validity was set forth in 1971 by Cronbach in a classic article titled "Test Validation." Cronbach described test validation as a process in which evidence is collected by the developer of a test to support the types of inferences that may be appropriately drawn from test scores. One immediately notices in Cronbach's definition that the emphasis of validity is not on the instrument itself; rather, the emphasis is upon the interpretation of the scores yielded by a test. Since Cronbach's first definition the psychological basis for valid interpretations of test scores has grown ever more important. More recently, Cronbach (1988) addressed the changing nature of validity by describing it as a concept to be viewed from varying perspectives.

Of course, Cronbach's psychologically-based definition of validity is in contrast to Garrett's (1937) traditional wording in which he described validity simply as ". . . the fidelity with which [a test] measures what it purports to measure" (p. 324). Although Garrett's definition is cited by some persons as the definition of validity, it is clearly much more limited than Cronbach's and reflects a difference in emphasis. Item writers should adopt Cronbach's more sophisticated view of validity.

One may look at validity from varying perspectives by examining the following definitions by other researchers.

Anastasi (1997): "The validity of a test concerns *what* the test measures and how well it does so" (p. 139).

Messick (1988): "Validity is an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores. As such, validity is an inductive summary of both the adequacy of existing evidence for and the appropriateness of potential consequences of test interpretation and use" (pp. 33-34).

Sax (1989): "Validity is defined as the extent to which measurements are useful in making decisions relevant to a given purpose" (p. 292).

Mehrens and Lehmann (1987): "Validity can be best defined as the extent to which certain inferences can *be made accurately from*—and certain actions should be based on—test scores or other measurement" (p. 265).

A Complete Definition of Validity

From these citations, the reader can easily recognize the significance of validity to the process of measurement in general and to constructing test items in particular. Let us consider, however, a fuller description of validity. Citing the AERA/APA/NCME *Standards* (1985):

> *Validity is the most important consideration in test development. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is a process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence*

> *supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself. (p. 9)*

Given the importance of the concept of validity to constructing good test items, it is worthwhile to closely examine the points made in this description. The first point to notice is that this description refers to *inferences* that are made rather than to direct measurements. This point should be unmistakably clear since it is consistent with the definition of a test item presented earlier and was discussed thoroughly in the section explaining the purposes for test items in Chapter 2.

Note especially in the description of validity that a given instrument is not itself validated; rather, it is the interpretation of the test scores that has meaning. Evidence for a particular interpretation of scores is evidence for validity. Thus, test validation is the process of gathering evidence for a specific interpretation of the scores yielded by a given test.

This important aspect of validity is not widely appreciated. It is a common misconception that validity is a particular phenomenon whose presence in a test may be concretely evaluated. One often hears exclamations that a given test is "valid" or "not valid." Such pronouncements are not credible; they reflect neither the focus nor the complexity of validity.

Further, test validation is a *process* of gathering evidence. As is noted in the description, there are many different methods for gauging and documenting evidence that justify particular inferences. Some methods for gathering evidence may corroborate data gathered by other methods. Conversely, some evidence may dilute the effect of or negate other types of support for specific inferences. And, some evidence may be appropriate for certain kinds of inferences but not for other inferences.

Actually, the evidence gathered establishes the kinds of inferences that may be appropriately made. For example, suppose a given set of test items is designed to assess verbal reasoning. If sufficient evidence supports the interpretation that a low score on the set of items means an examinee is low in the construct of verbal reasoning, then such an inference may be quite proper. However, it would not be correct to infer from the same low score that the examinee is also low in the construct of emotional maturity (or, for that matter, any other construct or ability). The evidence gathered only substantiates one type of inference, in this case the inference for an examinee's verbal reasoning ability.

Of course, it may be entirely possible to gather many types of evidence for valid interpretations of test scores, each directed at substantiating a different inference. Each inference, however, that leads to a conclusion about an examinee's performance must have its own supporting evidence.

Additionally, it is important to realize that validity is not like an on-or-off switch, but is expressed in degrees. A large amount of evidence may support a high degree of validity for certain inferences, a moderate amount of evidence supports inferences only moderately, and little evidence supports inferences only weakly.

Finally, the description of validity specifies that it is a *unitary concept.* Considering validity as a unitary concept means that there are not different "types" of validity. What had formerly beent thought of as *construct validity*, *content validity,* and *criterion-related validity*, each with independent criteria (cf. AERA, 1955; APA, 1966; APA/AERA/NCME, 1954), are now considered to be merely convenient categories of evidence for a single notion of validity. As is noted in the most recent edition of the AERA/APA/NCME *Standards* (1985), "The use of category labels does not imply that there are distinct types of validity . . ."(p. 9). Validity is a single notion.

As a unitary concept, validity may include several different types of evidence. The current conception refers to *construct-related evidence, content-related evidence,* and *criterion-related evidence.* These categories may be further delineated for convenience, but they are not in themselves different kinds of validity. In optimal circumstances, evidentiary support for validity is gathered from all three categories.

Construct-Related Evidence for Validity

From the item writer's point of view, evidence that validity is related to psychological constructs is especially important. The reader will recall from Chapter 2 that a psychological construct is a pattern of behavior consistently observed over a period of time (as, for example, reading ability), and that it is assumed that such traits can be indirectly assessed by test items. Many theoreticians believe that the notion of construct-related evidence is so intrinsic to making mental measurements by test items that all forms of evidence for validity actually fall under the generalized rubric of construct validation (Angoff, 1988; Guion, 1977; Messick, 1975, 1980, 1988; Tenopyr, 1977).

Other theoreticians, however, argue that actually establishing construct-related evidence for validity is difficult, if not impossible. For example, Ebel

(1983) asserts that while "the process of construct validation is intriguing . . . the product is seldom decisive . . . . This is a neat conception in the abstract, but it has turned out not to be very practical" (p. 10). Further, Ebel claims that most measures of school achievement and employee or professional competence assess skills that can be operationally defined and that these are not the kind of latent traits that Cronbach and Meehl (1955) had in mind when they defined construct validation originally. Therefore, Ebel and Frisbie (1986) argue, such tests

> *. . . should not require any special construct evidence of validity for the test user to make appropriate or meaningful inferences on the basis of the scores derived from them. Validity evidence is incorporated in the test-development process by rational statements about what abilities are measured and why the tasks are appropriate for measuring those abilities. (pp. 96-97)*

As one can see, considerable diversity of opinion exists about the practicality of establishing construct-related evidence for validity. Wainer and Braun (1988) and Mitchell (1986) document many important aspects of this controversy, and the interested reader is referred to these authors for a more complete analysis. Regardless, the importance of construct-related evidence for validity, even if only theoretically conceived, is not disputed.

THE RELATIONSHIP BETWEEN CONSTRUCTING TEST ITEMS AND VALIDITY

Abetting Validity in the Item Construction Process

It is important to note that even a careful, well-planned item-construction process does not constitute evidence for validity *per se.* Describing the steps taken to ensure that standards of quality are present in the items is important for producing good items and tests, but such a description is not direct evidence for validity. The procedures used during item construction do not authenticate a particular inference, nor do they offer direct proof that evidence for validity may eventually be garnered. Validity is concerned with the veracity of interpretations for test scores, not with how tests are constructed.

A rough analogy may be made to judging a musical recital. Imagine that a fine violinist is playing for a panel of expert judges. The judges attend to the sounds produced by the violinist, not to the steps taken by the violin maker to manufacture a quality instrument. If the judges feel moved by the music, this is akin to evidence for validity. The judges are more likely to be so moved if the violinist plays on a well-constructed instrument, but the construction process is not in itself grounds for the judges to be stirred to emotion. To be sure, the violinist will more likely produce music with the desired effect if he or she plays on a well-crafted violin rather than on an instrument whose characteristics are poor or unknown. So, too, constructing test items according to accepted standards of quality will enhance the likelihood that substantive evidence for valid test interpretations may be eventually gathered.

From this discussion one can garner a sense that the relationship between constructing test items and validity is both complex and important. Clearly, good items make valid interpretations for a test's scores possible, even though the procedure used to imbue quality into the items is not itself direct evidence for inferential interpretations.

Documenting the Item-Construction Steps

Documenting the steps taken to produce test items is necessary for two reasons. First, as with any endeavor in either the social sciences or the physical sciences, the work should be documented so that a knowledgeable person can replicate the task. One should anticipate that if the documented procedures are followed by another scientist, results similar (within chance fluctuations) to the original would be obtained. Within the context of constructing test items, this means that if an informed writer of test items followed accepted criteria for constructing good items and carefully documented the procedures he or she used, another informed item writer could replicate the procedures and expect items of about equal merit.

A second reason it is necessary to write down the steps used in the item-construction process is that the description itself will likely be of enormous assistance in determining whether a specific interpretation for a test's scores is valid. This fact is immediately apparent when one realizes that items for a given test are constructed with a specific purpose in mind and that this same test's scores have a particular interpretation. The more clearly articulated and understood are the purposes for the test items, the better one is able to gauge whether a specific interpretation for the scores is warranted.

CONDITIONS FOR ITEMS TO CONTRIBUTE TO VALIDITY

In addition to ensuring that a well-considered process is employed in constructing test items, certain conditions should be met for there to be eventual evidence for validity. These conditions are: 1) a well-defined purpose for the test, including precise delineation of the test's content, 2) a set of specifications for items consistent with the test's purpose and content, and 3) a defensible methodology for reviewing the congruence between the test items and their specifications. Each of these conditions should be carefully considered when seeking evidence for validity, especially when such evidence is content-related. They are conditions for constructing test items which set the stage for producing items that will meet the criteria for merit described in Chapter 2.

Defining a Test's Purpose

The first condition, clearly defining a purpose for the test and identifying the content, is extremely important. Obviously, one should eschew constructing items for tests for which only an ill-defined purpose has been established or one for which the content is not specifically identified. Although such a comment may appear so evident that it should scarcely warrant stating, Haertel and Calfee (1983) report that even this basic condition is routinely overlooked in test construction.

In addition to the obvious need for an articulated purpose for every test, the AERA/APA/NCME *Standards* (1985) dictate that test developers should compile evidence for the need for distinct information *prior to* publishing a particular test as well as specify the content intended to be sampled by the items. Without such clarity of content, the interpretation of a test's scores is necessarily restricted only to the performance on the specific items and cannot be generalized. Since most tests seek interpretations beyond just the performance on the specific items to larger content domains and to psychological constructs, a restricted interpretation would clearly be a disadvantage for a given test.

Further, the consequences of constructing tests for which there is no clearly defined purpose or domain of content have become increasingly serious in recent years. Litigation involving the denial of a property right, such as a high school diploma, on the basis of a low test score frequently includes discussions of ill-conceived purposes and inadequately specified content (Yalow and Popham, 1983).

A clearly defined purpose and domain of content are important for all tests, whether criterion-referenced, norm-referenced, or some other type. A common misconception is that since norm-referenced tests are typically geared to assess an examinee's relative performance on open-ended achievement, a precise definition of the intended content is unnecessary. Nothing could be further from the truth. Clearly defining the content to be tested is important in any kind of measurement, and is not a distinction between criterion-referenced measurement and norm-referenced measurement. In fact, recognizing the significance of elucidating the content antedates even the introduction of the term *criterion-referenced testing* (Ebel, 1962a; Flanagan, 1962; Nitko, 1984b).

Devising Specifications for Tests and Items

The second condition of item preparation for there to be eventual evidence for test validity is devising specifications for tests and items consistent with the test's purpose. This condition is important yet deceptively difficult to accomplish.

Regrettably, preparing specifications for tests or items is all too often ignored in test-item construction. One carefully done study of eleven widely used, criterion-referenced tests—all produced by commercial test publishers—revealed than none of the tests' developers used domain specifications when preparing the tests' items (Hambleton & Eignor, 1978). This is a sad commentary on the lack of care taken by many test developers. One hopes that with proper information, this omission will decrease for future test-production efforts. Two later sections of this chapter describe strategies for writing test specifications and item specifications.

Content-Related Evidence for Validity

The third condition for constructing test items is mainly relevant to amassing evidence for validity that is content-related. Throughout this chapter it has been emphasized that following the seven criteria for good items explained in Chapter 2 and the conditions just described does not in itself provide evidence for validity. Such evidence must be garnered through a validation study. A validation study, like all research efforts, should be conducted according to a rigorous methodology so that other researchers can reproduce the methodology and approximate the same results. Later, in Chapter 8, information is given on how to conduct a content-validation study as well as suggestions for gathering other types of validity evidence. If one follows the seven criteria for good items and the first two conditions for validity, there is a high likelihood that this third condition—a validation study—will yield a high degree of the desired evidence.

INITIAL CONSIDERATIONS WHEN SELECTING CONTENT FOR ITEMS

The Relationship Between Tests and Curriculum

Although it is emphasized throughout this book that test items are measures that reflect psychological constructs, it is tempting to unthinkingly adopt the proposition that tests automatically reflect curriculum. Accordingly, this faulty logic may continue, the subject content for a given test's items will always directly emanate from the same primary sources as the curriculum, as for example, textbooks and other program materials. This simplistic postulate can lead one to imagine that merely opening a textbook or other curriculum material will be sufficient when considering subject matter for test items. However, the relation between tests and curriculum materials or programs is not so simple.

Merely opening a textbook and writing items is an uninformed approach to determining the content of test items that can lead to a gross mismatch between what is measured and what is intended for measurement. Further, when this mismatch is not recognized, the errors in score interpretation can be chronic. It is necessary, then, to take a more informed approach to determining the content for test items, an approach that considers not only the precise subject matter, but the goals for the curriculum as well as the purposes for measurement. In this section we will examine the complex relationship between curriculum and assessment, and in several subsequent sections we will develop strategies to use this informed background to advantage in constructing good test items.

The reader will recall from earlier discussions in this book that tests are indirect measures of psychological constructs. Accordingly, test items do not simply restate curriculum facts. To produce good tests and good test items, one must be aware of how learning experiences impact psychological processes in students. The place to start such an inquiry is knowledge of the curriculum's basis and not merely its content. A few leading questions may direct one to the appropriate focus: Is the purpose of an instructional experience to communicate particular facts, or to imbue an appreciation for the significance of the facts? Or, is it to stimulate in the student a willing desire to learn? Or, is it for some other purpose?

Obviously, the same curriculum as well as a nearly identical lesson plan could serve any, or even all, of the purposes for a curriculum mentioned above. But identical test items would not ordinarily be used for assessing such diverse goals. More usually, different test items—with different content—would be used to assess the various curricular goals. Without identifying the purpose or the basis for the curriculum it would not be possible to ascertain the appropriate content for items. It is apparent, then, that the item writer must begin the search for appropriate content for test items, not from the textbook that may have been used to teach the curriculum, but from an awareness of the goals of the curriculum.

Modern Goal Conception

Tyler (1949), in his imaginative and seminal work on establishing behavioral objectives, advocated that, "Tests must actually be based upon the objectives of instruction rather than simply sampling the content of instruction . . ." (p. 402). While many psychologists and others have advanced Tyler's work to more sophisticated levels, his essential point—that the adequacy of measurement by test items rests squarely upon those items reflecting clearly formulated instructional objectives—remains relevant to determining the content for test items.

Today, psychological understanding of cognitive learning has advanced to the point where it is possible to devise instructional goals in the language of cognition. These new interpretations seek to delineate educational goals in a way that permits subject content to fit in the context of an examinee's mental processes. A leading educational psychologist, R. E. Snow (1980), describes this enlightened view as follows.

> *Modem cognitive psychology now forces upon us a much richer conception of achievement than educational psychology heretofore embraced. The view of complex learning now extant emphasizes the organization, representation, and use of knowledge over the long haul, not just its short-term accumulation. . . . School achievement is no longer to be understood as simply the accretion of facts and content-specific skills; but, rather, a significant part of the learner's task is to continually assemble, reassemble, structure, and tune the*

> *cumulating body of knowledge into functional systems for use in thought, and in further learning. Thus, achievement is as much an organization function as it is an acquisition function,* (pp. 42-43)

Such complex considerations make the task of ascertaining the content for test items more challenging, but also more worthwhile. The task is more difficult because the item writer must adopt a wider perspective than merely considering subject content. The item writer must also contemplate how well a particular test item represents a specific psychological construct and how an examinee's response to the item may provide inferential clues to his or her ability level in that construct. Lohman and Ippel (1993) describe cognitive theory as a backdrop for understanding complex learning and knowledge acquisition. With this enlightened view, the task of selecting appropriate content for test items is more meaningful because well-constructed test items can significantly contribute to eventual evidence for valid interpretations of a test's scores.

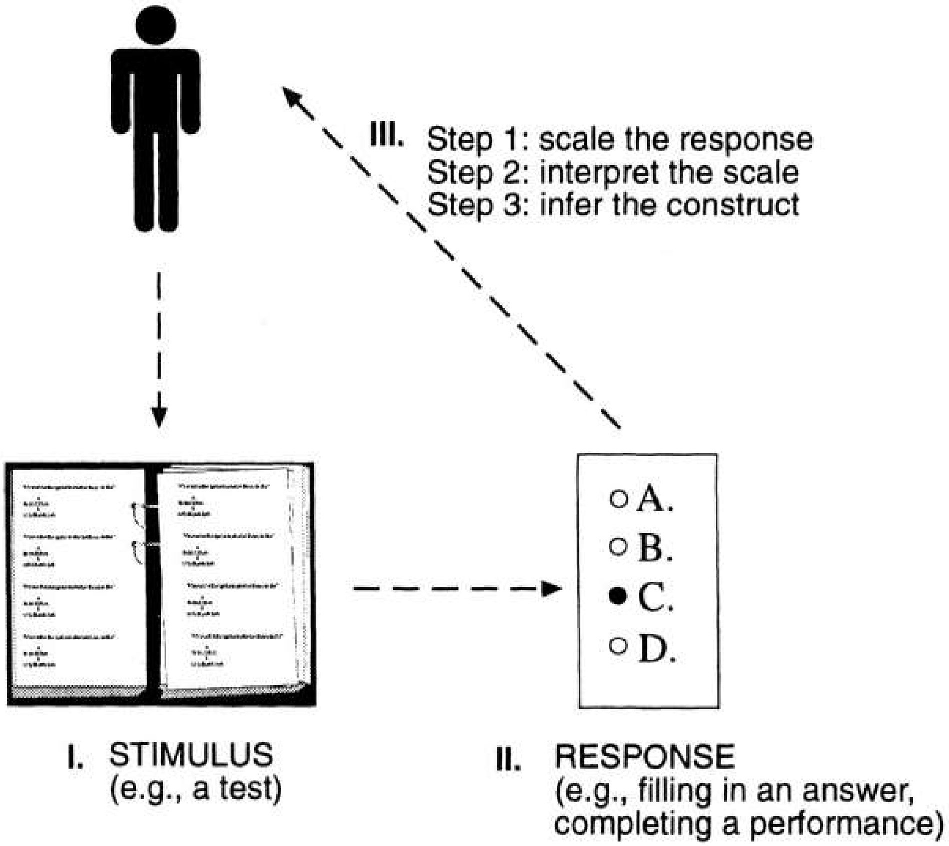
This point is emphasized in Figure 3.1. This figure depicts the "assessment triangle" inherent in all test items. By the triangle, one can see that the examinee is presented with a stimulus situation, namely, the test item (which can be of virtually any format, whether supply-type, constructed response, or some performance). The examinee, then, responds to this stimulus by answering the question (e.g., bubbling in the answer sheet, executing the performance). The examinee's response is then, first scaled (adapted to a numbering scheme), second, the scale is interpreted, and finally, an inference is made from this interpretation to the construct.

ACHIEVING CLARITY IN A TEST'S CONTENT

Importance of Specificity in Content

As has been emphasized throughout this book, a clear understanding of the content intended for a test is central to constructing good test items. Unfortunately, achieving clarity in articulating content is difficult and is often only superficially attempted. To merely report that a test is measuring language arts or quantitative skills or some other loosely defined content is to misunderstand the importance of specifying the test's content.

Figure 3.1 Representation of assessment triangle.


III. Step 1: scale the response
Step 2: interpret the scale
Step 3: infer the construct
○A.
○B.
●C.
○D.
I. STIMULUS
(e.g., a test)
II. RESPONSE
(e.g., filling in an answer, completing a performance)

To illustrate the point, consider the following example. Suppose a test developer wishes to construct a test designed to assess social studies. He or she communicates this to the item writer. While this is necessary to the item writer's work, it is insufficient information for the test item writer to do a good job. Too much is not stated. The number of items that could legitimately fall under the category "social studies" is limitless. The item writer could just as easily select the content from among myriad historical facts, or from the fields of geography, economics, political science, or even the social sciences. The measurement of a construct so inexactly defined can provide no meaningful assessment. Regretfully, this scenario is all too common.

Now, suppose the test developer specified to the item writer that only history can be included in the content of this social studies test. This limitation greatly enhances the specificity of content. The item writer's task is now more manageable but still not sufficiently clear to provide real focus to the items. If the test developer further delimited history to include one specific objective—for example, to recognize the chronology and significance of major events and movements in United States history—the clarity of content would be advanced. The test developer has articulated a clearly defined domain of content within which the item writer can focus his or her efforts. The probability of overall evidence for valid content-related interpretations of the test scores is similarly increased.

Examples of Clearly Defined Content

It is possible to describe the content with even more specificity than is provided by the objective cited above. The utility of further clarity will depend upon the purpose for the test. If the test is designed to be a criterion-referenced measure of particular subject content, then even further limitation of content is desirable. Table 3.1 provides an example of a very focused social studies content. Notice the precision of language, defining the content in terms of subject area, cluster, skill, and enabling subskill, including the annotations. This content description prescribes the item writer's task beneficially and is an excellent example of clearly stated test content.

Table 3.2 presents another good example of clarity of description for a test's content. This example is from a test of reading comprehension, a commonly tested area that is often only vaguely articulated. Here, real clarity is achieved and the item writer's task is evident.

The scenario described above, and augmented by the examples of good content descriptions in Tables 3.1 and 3.2, suggests that carefully crafted wording is worth the effort. The skilled item writer should work with the test developer in achieving this degree of understanding of a test's content.

Expert test developers such as Baker (1974), Millman (1974b), and Popham (1975; 1984) all advocate specificity in description of content as a way to improve understanding between the developer of a test and the item writer. These researchers describe their work variously as *amplified objectives, domain specifications,* or *test content specifications.* Some of these terms are less in fashion now and precise distinctions among these terms are not of substantive importance when planning for item construction since they all focus on precision in language to aid understanding so that test items may optimally reflect their objective. For our purposes, we consider all of these efforts to fall under the general rubric of "test content specifications." Such test content specification typically have three elements: 1) a

Table 3.1 Content Description for a Social Studies Skill

Subject Area: Social Studies

Cluster: History

Skill: Recognize the chronology and significance of major events and movements in United States history.

Enabling Subskills:

A. Identify and compare key institutions and participants[1] in major events and movements of United States history.

B. Identify the sequence of major events and movements[2] in United States history.

C. Describe the significance of major events and movements in United States history, including their causes and effects as well as their relationships to broader historical trends.

D. Identify technological developments and environmental changes[3] in United States history and relate them to historical events and movements.

E. Describe the principles and development of American Constitutional democracy and the significance of major Supreme Court decisions.

F. Describe the interaction among peoples of different national origins, races, and cultures and how such interaction has shaped American history.

[1] For example, public schools, daycare industry, New York Stock Exchange, Chicago Commodities Exchange, Congress; Thomas Jefferson, Susan B. Anthony, Carrie Nation, Franklin Roosevelt, Martin Luther King, Jr.

[2] For example, Revolutionary War, Louisiana Purchase, Lincoln-Douglas debates, Civil War, populist movement, woman suffrage, Prohibition, Great Depression, civil rights movement, first moon landing.

[3] For example, telephone, automobile, airplane, satellite communications, genetic engineering, acid rain, depletion of the ozone layer, deforestation of rain forests.

From *College BASE Guide to Test Content* (p. 16) by S. J. Osterlind, et al., 1989, Chicago, IL: Riverside. Reprinted by permission.

Table 3.2 Content Description for a Reading Comprehension Skill

Subject Area: English

Cluster: Reading and Literature

Skill: Read a literary text analytically, seeing relationships between form and content.

Enabling Subskills:

A. Identify and analyze common semantic features such as connotation and figures of speech.

B. Identify conventional literary genres, elements, and devices[1] and relate such formal elements to the content of the passage in which they are found.

C. Identify the tone, mood, and voice of a literary text through an analysis of its linguistic features and literary devices.

D. Identify the theme of a literary text and the ways it is embodied by formal elements.

[1] For example, sonnet, epic, lyric, conflict, setting, exposition, blank verse, couplet, point of view.

From *College BASE Guide to Test Content* (p. 3) by S. J. Osterlind, et al., 1989, Chicago, IL: Riverside. Reprinted by permission.

description of the content areas that are to be tested, 2) a statement of the objectives or mental processes to be assessed, and 3) a description of the relative importance of #1 and #2 to the overall test.

Establishing a Table of Test Content Specifications

One convenient way to establish a set of test content specifications is in terms of a table. The table should show the three basic elements of test content specifications: content, processes, and the importance of each. Such a table may be organized with processes across the top and content topics in the left column. Table 3.3 presents an example of using a two-way grid as a table of test content specifications.

Table 3.3 Sample Table of Content Specifications

Skill: Read critically by asking questions about a text, by recognizing assumptions, and by evaluating ideas.

Major Content Areas	**Intellectual Process***		
	Interpretive Reasoning	Strategic Reasoning	Adaptive Reasoning
Identify the literal meaning of a text and recall its details.	4	3	
Identify the main idea of a text and differentiate it from subordinate ideas.	2	4	
Summarize the rhetorical development or narrative sequence within a text.	4	4	1
Recognize the implicit assumptions and values that inform a text.	3		2
Assess the logical validity of the rhetorical development within a text.	3		1
Evaluate ideas in a text by their implications and relation- ships to ideas outside the text.	1	2	

* number of items

Note in this table, which gives test content specifications for a test of reading comprehension, that the left side depicts the content areas to be included on this test of critical reading skills. Across the top of the table, the intellectual processes are listed: Interpretive Reasoning, Strategic Reasoning, and Adaptive Reasoning. Also notice in the table that the relationship between the two dimensions is expressed as the number of items to be assigned to each category. Thus it can be seen that all of the elements of test content specifications are given in this table. Note also that not every cell is filled, indicating that no items are required for some areas.

Table 3.4 presents another example of a table of test content specifications. In this table, the content areas are listed along the top and the intellectual processes (here they are called "process objectives") are cited in the left column. And, the table indicates the relation between the two as the approximate number of items that are to be included in the final test assigned to each content area and process objective. The percentage of total items intended for each category is also given for each column and row. Also, note that this table includes in every cell very specific information about the content for the items. This narrowly focused content is common in tests that are criterion-referenced, and typical in many tests used for licensing and certification.

While it is often useful to organize test content specifications into a two-way table, there is a tendency to believe that every cell must be filled or a deficiency exists. This is not accurate. For some tests, classifications along a single dimension may be more appropriate. Alternatively, complex tests may require more dimensions and fuller descriptions of its characteristics. For example, various formats for the items can also be specified in a table of test content specifications. Table 3.5 displays a table of test content specifications that includes directions for item formats in addition to identifying the content and processes.

As can be seen in Table 3.5, the left side of the table consists of the major skill groups (e.g., "Capitalize Words," "Common Nouns"). Across the top of the table are the major cognitive levels of recall and application. Below these levels are the subordinate skills and the behavioral objectives that provide the criteria for the attainment of the subordinate skills.

Sometimes the instructions for item format are displayed in item specifications, rather than in test specifications. There is no particular advantage to displaying item formats in test specifications unless there are no item specifications offered for a particular test. In these cases, the prescription of an item format is desirable.

Alternative Ways to Present Test Content Specifications

While tables of specifications such as those displayed in Tables 3.3, 3.4, and 3.5 are commonly used for presenting test content and other relevant information, and may be adequate for a particular instrument, they are not the only possible manner of organizing test specifications. Complex sets of specifications are not uncommon. In fact, Ebel and Frisbie (1986) stipulate that the firmest basis for a good set of test content specifications should include fully seven elements, indicating the following:

- formats of test items to be used,
- number of items of each format,
- kinds of tasks the items will present,
- number of tasks of each kind,
- areas of content to be sampled,
- number of items in each area, and
- level and distribution of item difficulty.

Systematic attention to all seven of these elements will facilitate the development of comprehensive test content specifications. Clearly, such thorough test content specifications will give a tight focus to the item writer's task. But regardless of whether the test content specifications are complex or comparatively simple, they should address at least the three basic elements (viz., content, processes, and importance of each).

The Test Developer's Responsibility for Preparing Test Content Specifications

Preparing test content specifications is the responsibility of the test developer. Such preliminary organization precedes the work of actually writing test items. The item writer will use this information to carry out the intentions of the test developer. Frequently, however, the work of the test developer and the item writer is carried out by the same person or group of persons. This dual role can be an asset to good test development because the goals of the test may be more precisely understood by the item writer if he

[Text continues on page 83.]

Table 3.4 Sample of Test Content Specifications for a Biology Test.

	Process Objectives		
CONTENT AREAS	Identifies Specific Facts 20%		Recognizes Terms and Vocabulary 30%
A. Nutrition, 40%	Nutrients Vitamins Enzymes Metabolism Oxidation	Incomplete Protein Complete Protein Amino Acids Glycogen Carbohydrate	Nutrients Essential to Health Good Sources of Food Nutrients Parts of Digestive System Process of Digestion of Each Nutrient Sources of Information About Foods and Nutrition
B. Communicable Disease, 40%	4 or 5 items		7 or 8 items
	Immunity Virus Carrier Antibodies Incubation Period	Epidemic Pathogenic Endemic Protozoa	Common Communicable Diseases Incidence of Various Diseases Methods of Spreading Disease Types of Immunization Symptoms of Common Communicable Diseases
C. Noncommunicable Diseases, 20%	4 or 5 items		7 or 8 items
	Goiter Deficiency Diseases Diabetes Cardiovascular Diseases Caries		Specific Diseases Caused by Lack of Vitamins Specific Disorders Resulting from Imbalance in Hormones Incidence of Noncommunicable Diseases Common Noncommunicable Diseases of Adolescents and Young Adults
Number of Items	3 or 4 items		2 or 3 items
	18		12

From *Measurement and Evaluation in Psychology and Education*, 4th ed. (pp. 18-21) by R. L. Thorndike, 1977, New York: Wiley. Adapted by permission.

Table 3.4 *(continued)*

	Process Objectives		
Identifies Principles, Concepts, and Generalizations 30%	Evaluates Health Information and Advertisements 10%	Applies Principles and Generalizations to Novel Situations 10%	Number of Items
Bases of Well Balanced Diet Enzyme Reactions Transfer of Materials Between Cells Cell Metabolism Functions of Nutrients in Body	Analyzes Food and Diet Advertisements Interprets Labels on Foods Identifies Good Sources of Information About Foods and Diets	Identifies Well Balanced Diet Computes Calories Needed for Weight-Gaining or Weight-Losing Diet Predicts Consequences of Changes in Enzymes on Digestive System Identifies Services and Protection Provided by the Federal Food and Drug Act	
7 or 8 items	2 or 3 items	2 or 3 items	24
Basic Principles Underlying Control of Disease Actions of Antibiotics Body Defenses Against Disease Inmune Reactions in Body	Distinguishes Between Adequate and Inadequate Evidence for Medicines Identifies Misleading Advertisements for Medications	Recognizes Conditions Likely to Result in Increase of Communicable Disease Identifies Appropriate Methods for Sterilizing Objects Gives Appropriate Reasons for Regulations, Processes, or Treatments	
7 or 8 items	2 or 3 items	2 or 3 items	24
Pressure Within Cardiovascular System Control of Diabetes Inheritance of Abnormal Conditions Abnormal Growth of Cells	Identifies Errors or Misleading Information in Health Material Identifies Appropriate Source of nformation for Health Problems	Predicts Consequences of Changes in Secretion of Certain Hormones Predicts Probability of Inheriting Abnormal Conditions	
3 or 4 items	1 or 2 items	1 or 2 items	12
18	6	6	Total number of items 60

Table 3.5 Sample Table of Test Content Specifications that Includes Information on Item Formats

Recall (Knowledge and Comprehension)				
Major Skill Group	*Subordinate Skills*	*Behavioral Objectives*	*Item Format*[a]	*Number of Items*
Capitalize words (prerequisite skill)	1. State the rule for capitalizing words	1.1 From memory state the rule for capitalizing words.	W	1
		1.2 Select the rule for capitalizing words from a set of alternative rules.	S	1
Common nouns (prerequisite skill)	3. Define terms *person, place, thing,* and *idea*	3.1 From memory, define the terms *person, place, thing,* and *idea.*	W	4
		3.2 Match the terms *person, place, thing,* and *idea* with their definitions.	S	4
	5. Define the term *noun*	5.1 From memory, define the term *noun.*	W	1
		5.2 Given the definition for term *noun,* identify it as such.	W/S	1

From *Measuring and Evaluating School Learning* (pp 84-85) by Carey, L. M., 1988, Boston: Allyn and Bacon. Adapted by permission.

Table 3.5 *(continued)*

	Application		
Subordinate Skills	*Behavioral Objectives*	*Item Format*	*Number of Items*
2. Capitalize any word	2.1 Given a list of several words, select the letter in the words that should be capitalized.	S	5
	2.2 Given some words that are properly capitalized and some that are improperly capitalized, select those that are properly capitalized.	S	5
4. Discriminate words that represent persons, places, things, and ideas from lists of words containing these and other concepts	4.1 Given a list of words that contains persons, places, things, and ideas, classify each word into the appropriate category	S	12
6. Classify words that are common nouns	6.1 Given a list of words containing common nouns and other parts of speech, select the common nouns.	S	15
7. Give several examples of common nouns	7.1 List several words that refer to persons.	W	2
	7.2 List several words that refer to places.	W	2
	7.3 List several words that refer to things.	W	2
	7.4 List several words that refer to ideas.	W	2

[a] Item format codes: W = Write response from memory; S = Select response from alternatives; W/S = Either write or select response.

or she is the same person who developed the goals originally. One caution to be heeded with this two-in-one role, however, is that it can be tempting to cut corners in the test development process by only visualizing the test content specifications rather than actually writing them out. This is a poor substitute for writing clear, well-thought-out test content specifications.

Although developing test content specifications may appear to the untutored item writer to be a laborious step that is tempting to skip, they are crucial to a skillful determination of the content for test items. They are simply too important to ignore.

MELDING COGNITIVE PROCESSING LEVELS WITH ITEM CONTENT

Identifying Levels of Cognitive Processing

Bloom's taxonomy is probably the most widely employed scheme for labeling and articulating levels of cognitive processes in test construction today, as it has been for the past two decades. It is used in variously modified versions by the developers of many popular tests. Bloom's original work in describing mental processing in the cognitive domain was seminal, bringing widespread attention to the notion of classifying psychological processes into categories that could be exploited for measurement. The taxonomy has been elaborated upon and fitted with numerous examples of test items for each category in a later publication by Bloom, Hastings, and Madaus (1971). The taxonomy includes these primary categories, described in Bloom's original wording:

> 1.00 KNOWLEDGE
>
> Knowledge, as defined here, involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting. For measurement purposes, the recall situation involves little more than bringing to mind the appropriate material. Although some alteration of the material may be required, this is a relatively minor part of the task. The knowledge objectives emphasize most the psychological processes of remembering. The process of relating is also involved in that a knowledge test situation requires the organization and reorganization of a problem such that

it will furnish the appropriate signals and cues for the information and knowledge of the individual possesses. To use an analogy, if one thinks of the mind as a file, the problem in a knowledge test situation is that of finding in the problem or task the appropriate signals, cues, and clues which will most effectively bring out whatever knowledge is filed or stored.

2.00 COMPREHENSION

This represents the lowest level of understanding. It refers to a type of understanding or apprehension such that the individual knows what is being communicated and can make use of the material or idea being communicated without necessarily relating it to other material or seeing its fullest implications.

3.00 APPLICATION

The use of abstractions in particular and concrete situations. The abstractions may be in the form of general ideas, rules of procedures, or generalized methods. The abstractions may also be technical principles, ideas, and theories which must be remembered and applied.

- Application to the phenomena discussed in one paper of the scientific terms or concepts used in other papers.
- The ability to predict the probable effect of a change in a factor on a biological situation previously at equilibrium.

4.00 ANALYSIS

The breakdown of a communication into its constituent elements or parts such that relative hierarchy of ideas is made clear and/or the relations between the ideas expressed are made explicit. Such analyses are intended to clarify the communication, to indicate how the communication is organized, and the way in which it manages to convey its effects, as well as its basis and arrangement.

5.00 SYNTHESIS

The putting together of elements and parts so as to form a whole. This involves the process of working with pieces, parts, elements, etc., and arranging and combining them in such a way as to constitute a pattern or structure not clearly there before.

6.00 EVALUATION

Judgments about the value of material and methods for given purposes. Quantitative and qualitative judgments about the extent to which material and methods satisfy criteria. Use of a standard of appraisal. The criteria may be those determined by the student or those which are given to him.

Difficulty of Using Bloom's Taxonomy to Write Items

Regardless of its widespread use, Bloom's taxonomy of educational objectives is often constrictive to test developers. This is so because developing test items which conform to the language of the taxonomy results in too many items being labeled in the knowledge dimension, the lowest-level category. With care, some items may be written to the comprehension or even the application category, but by Bloom's descriptions few items can be constructed to assess processes at the high-end taxonomy levels of analysis, synthesis, or evaluation.

This deficit has become increasingly apparent in recent years with the closer scrutiny by modern measurement experts of items that tap complex cognitive processing. Concomitantly, as test developers and item writers place more importance on the language of cognition, there is a growing sensitivity to the importance of defining the levels of cognitive processing and identifying how particular test items assess mental processing. No longer is it adequate to assign items cursorily to process categories and pay little heed to the consequences. Today, the careful item writer must be precise in the definitions of cognitive levels adopted for use, and must consider carefully the precision with which particular test items may tap specific levels of mental processing.

Yet another problem arises with Bloom's taxonomic scheme: there is an inherent difficulty in validating the properties of the levels within the taxonomy. Madaus, Woods, and Nuttall (1973) claim that Bloom's scheme has no structural hierarchy beyond what can be explained by a general intelligence, or "g" factor. And, Seddon (1978), after reviewing the relevant literature, maintains that no one has been able to demonstrate the veracity of Bloom's levels. An extreme position is advocated by Blumberg, Alschuler, and Rezmovic (1982), who state that Bloom's scheme should not be used for test development at all until the "significance of taxonomic levels has been established . . ." (p. 6). These researchers give us a clearer understanding of the merits and limitations of using Bloom's taxonomy of the cognitive domain for developing test items.

Over the years, theoreticians and researchers have sought to develop Bloom's idea of making a taxonomy of educational objectives and have had new and important perspectives. For example, Ebel (1972), Thorndike and Hagen (1977), Hannah and Michaelis (1977), and others have contributed their own taxonomies of mental processing for the cognitive domain. Most of these taxonomies have attempted to provide test developers and item writers with more felicitous descriptions of the cognitive domain than Bloom's.

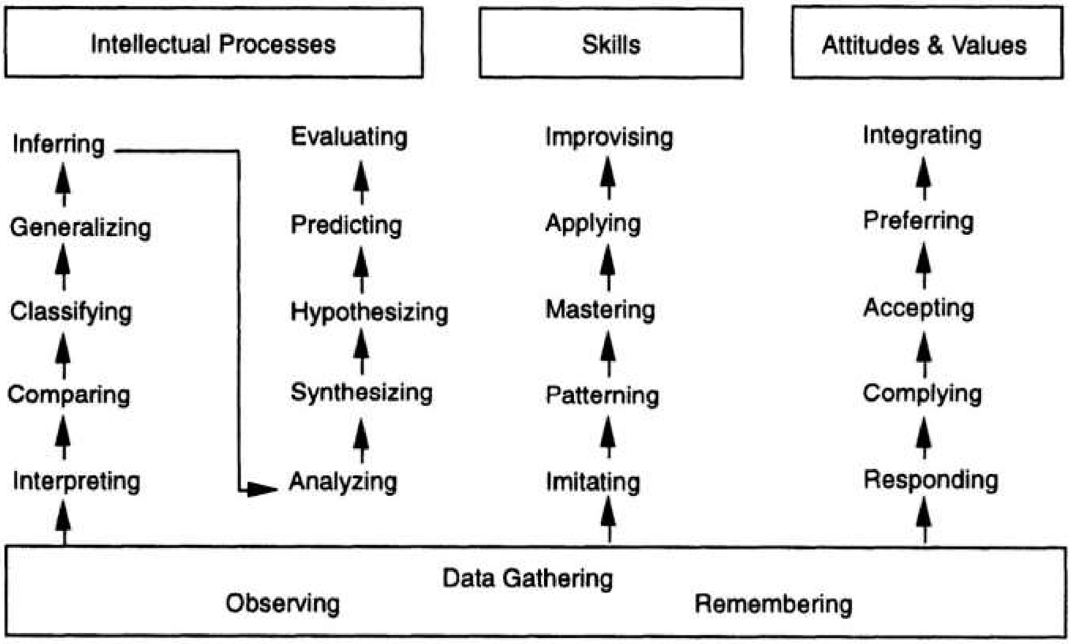
The Framework for Instructional Objectives Taxonomy

Of the theoreticians mentioned above, the taxonomy developed by Hannah and Michaelis (1977) has the most potential for aiding construction of test items. The Hannah and Michaelis system, called the *Framework for Instructional Objectives* (FIO), was developed to give educators the generic vocabulary and common perspective needed to integrate instruction with evaluation by stressing the interrelated nature of these activities. By providing a detailed framework for writing instructional objectives, FIO aids the design of objectives-based assessment. The primary function of FIO is utilitarian. It does not break any new theoretical ground.

FIO is divided into four domains, with each domain separated into levels. These are illustrated in Table 3.6. Understood as a hierarchy, the levels—called behaviors—define the domains. The first and primary domain, data gathering, consists of observing and remembering. These are prerequisite to the more complex learning processes. This domain underpins the other three domains so that observing and remembering act as the understood first and second levels of the behaviors in each domain. Roughly speaking, FIO uses the designations *Intellectual Processes* for cognitive behavior, *Skills* for psychomotor behavior, and *Values and Attitudes* for affective behavior. Of course, there is overlap, and some of the intellectual processes of the cognitive domain would be requisite to the values and skills domains. The levels of each domain are arranged in a hierarchical order according to the criteria which defined the domain's behavior: complexity (Intellectual Processes), independence (Skills), and integration (Attitudes and Values). Table 3.6 illustrates the relationship of the levels of behavior inside the domains.

Of FIO's three major domains, only Intellectual Processes clearly lends itself to evaluation through the kinds of test-item formats discussed in this book (viz., selected-response and constructed-response). Further, Intellectual Processes is also the domain which relates directly to the traditional

Table 3.6 Framework for Instructional Objectives



From *A Comprehensive Framework for Instructional Objectives* (p. 16) by L.S. Hannah, and J. U. Michaelis, 1977, Reading, MA: Addison-Wesley. Reprinted by permission.

school curriculum: English, mathematics, science, social studies, and the like. Therefore, this domain may be the most widely applied in the process of constructing test items.

FIO is useful in constructing test items because it provides an internally consistent framework from which one can discuss content-related evidence and construet-related evidence for validity. To appreciate this aspect of FIO, recall that it was mentioned earlier that it is difficult to prepare test items consistent with Bloom's taxonomy in any category beyond the lowest one or two (viz., knowledge and comprehension). The FIO classification scheme is more flexible because it does not contain a formal definition of the domain of knowledge as such; rather, each level contains lists of suitable objects and conditions for the desired behaviors, and these lists provide the information usually contained in classifications of the domain of knowledge. For instance, an overt behavior for the level of Interpreting is "restates," and a suitable object for this behavior is "main ideas."

Similarly, an overt behavior for the level of Classifying is "names," and a fitting object for this behavior is "objects." This level of detail is more appropriate to actual construction of items than to planning for test specifications. Applying these behavioral descriptions to item construction is discussed fully in Chapter 4, with a complete listing of the categories and their concomitant behavior provided. Now, however, the point to notice is how the FIO may be used when preparing test content specifications.

TEST ITEM SPECIFICATIONS

Test item specifications are a specialized kind of technical writing used in developing a set of items. Just as test content specifications describe the content and intellectual processes for an entire test, item specifications give directions for preparing particular items. However, the similarity between test content specifications and item specifications is only general. Test item specifications are not merely a more specific version of test content specifications. They differ in purpose, scope, and function. Test item specifications are formal, systematized directions from a test developer to the item writer that seek to put the test content specifications into action. They may include such information as eligible item formats, kinds of directions, limits for the stem, characteristics of the response alternatives, as well as features for the correct response and distractors.

Test item specifications can be brief, or they may need to be lengthy if the test developer wishes to convey a lot of information to the item writer. The purposes of the test will, in great measure, dictate the amount of information needed to describe item specifications.

An example of an item specification is given in Table 3.7 below. Obviously, when preparing item specifications, stating the goal, objective, skill, or standard for assessment is necessary, and it is usually given first. In the sample item specification, the subject for assessment is science, and it is further limited to laboratory and field work. The objective to be assessed by items in Table 3.7 is as follows: Recognize the role of observation and experimentation in the development of scientific theories.

Note in the sample item specifications, that the content for potential items has been broken into three types, Item Types A, B, and C. Collectively, the descriptions for the three item types define the entire skill. By comparing

[Text continues on page 95.]

Table 3.7 Sample of Test Item Specifications for a Science Test

Subject Area: Science

Cluster: Laboratory and Field Work

Skill: Recognize the role of observation and experimentation in the development of scientific theories.

Item Format: Multiple Choice, four responses.

Item Type A: Isolate and define a scientific problem or area for scientific study.

1) Stimulus Characteristics:
   a. The stimulus will direct the student to identify a statement that best defines a scientific problem or topic for investigation.
   b. The stimulus will provide a thorough description of observational data from which the student will discern the area to be investigated.
   c. The material should neither be so well known that the student is likely to be familiar with actual scientific work in the area nor so technical that specialized knowledge is required.

2) Response Options:
   a. The correct answer will be a sentence or brief paragraph that presents an accurate and clearly defined statement of the scientific problem or area of study.
   b. Distractors will be sentences or brief paragraphs that fail to define the scientific problem or area of study accurately, and which may be incorrectly or too broadly stated or may describe material that is irrelevant or outside the scope of the observation.

Table 3.7 *(continued)*

Sample Item: A researcher studying the use of radiation to slow the spoilage of vegetables harvested four dozen tomatoes of a variety that had been grown under identical conditions and were approximately the same size and weight. After randomly selecting them for placement in sterile racks containing a dozen each, she subjected two racks to a fixed amount of radiation. Then she placed one irradiated rack and one that had not been irradiated in refrigerators which maintained the same temperature. She stored the other two racks, one of which had been irradiated, at room temperature. She checked all four racks every six hours for signs of spoilage. When she found signs of spoilage in at least three of the tomatoes in a rack, she recorded the elapsed time and disposed of the tomatoes in that rack. The researcher found that the irradiated tomatoes spoiled at the same rate as those that were not irradiated. Repetition of the experiment gave the same results.

Which variable would it be best for the researcher to alter in her follow-up experiment?

A. amount of radiation

B. storage temperature

● C. number of tomatoes

D. size of tomatoes

Item Type B: Recognize the principal elements in an experimental design, including the hypothesis, independent and dependent variables, and controls.

1) Stimulus Characteristics:

a. The stimulus will direct the student to identify a hypothesis control, dependent variable, independent variable, result, or conclusion in a specified experiment design.

b. The material should neither be so well known that the student is likely to be familiar with actual

Table 3.7 *(continued)*

scientific work in the area nor so technical that specialized knowledge is required.

c. Although the research may be fictitious, the situation will be realistic and the goal practical.

2) Response Options:

a. The correct answer will be a word, phrase, sentence, or brief paragraph describing the experimental element called for in the stimulus.

b. Distractors will be words, phrases, sentences, or brief paragraphs describing other aspects of the experiment or, in the case of a hypothesis, stating overly general, overly specific, or simply erroneous summaries of what is being investigated in the experiment.

Sample Item:

Science

In the nineteenth century, Louis Pasteur performed an experiment in which he bent the necks of flasks into "S" shapes, leaving their ends opened. Then he boiled broth in the flasks to force air out and kill any microbes inside. After the flasks cooled, he left some of them upright for observation. Before setting aside others to observe, he tilted them so that the broth moved up into the bent necks and then back into the flasks. After the flasks had been prepared, he watched them for signs of microbial growth.

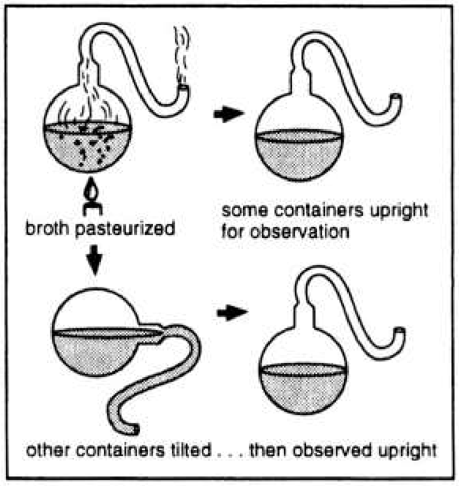

broth pasteurized
some containers upright for observation
other containers tilted . . . then observed upright

Table 3.7 *(continued)*

Which hypothesis was Pasteur testing in this experiment?

A. Flasks with bent necks would cause microbes to grow in the broth.

B. Cooling broth in the flasks would cause microbes to grow in the broth.

● C. Heating broth in the flasks and then cooling it would cause microbes to grow in the broth.

D. Contact of the broth with something in the necks of the flasks would cause microbes to grow in the broth.

Item Type C: Evaluate an experimental design by analyzing its ability to test the hypothesis, identifying weaknesses and improvements, and discerning inherent limitations and assumptions.

1) Stimulus Characteristics:

   a. The stimulus will direct the student to identify a flaw, improvement, limitation, or assumptions in a specified experimental design.

   b. The material should neither be so well known that the student is likely to be familiar with actual scientific work in the area nor so technical that specialized knowledge is required.

   c. Although the research may be fictitious, the situation will be realistic and the goal practical.

2) Response Options:

   a. The correct answer will be a word, phrase, sentence, or brief paragraph describing the flaw, improvement, limitation, or assumption called for in the stimulus.

Table 3.7 *(continued)*

b. Distractors will be words, phrases, sentences, or brief paragraphs describing flaws not present in the experiment, aspects of the experiment that are **not** flaws, changes that do **not** improve the experiment, assumptions that do **not** apply to the experiment or its conclusions.

Sample Item: **Read the passage below and answer the question that follows.**

A researcher studying a species of hawk noticed that an increasing number of their eggs were not hatching. He suspected that the problem was related to heavy use of a new pesticide in the area. To investigate the matter further, he acquired three hawk eggs immediately after they had been laid and placed them in an incubator. Next, he diluted a sample of the pesticide, which was not water soluble, in alcohol. He then injected one egg with the solution of pesticide and alcohol and a second egg only with alcohol. He sealed the injection sites in both eggs with wax. The third egg received no injection. He placed all three eggs back in the incubator for observation.

109. Which would most clearly improve the experiment?

A. sealing the injection sites with tape rather than wax

● B. using more eggs for each of the three conditions

C. soaking the eggs in pesticide rather than injecting it

D. experimenting with newborn hawks rather than eggs

Table 3.8 Sample Test Item Specifications for Minimum Student Performance Standards in Computer Literacy

GRADE:	5	SUBJECT: Computer Literacy
STANDARD:	D	The student will recognize the impact of computer technology in society and the need for its ethical use.
SKILL:	22	Identify an example of a computer application in each of the following areas: home, school, and business.
CLARIFICATION OF SKILL:		The student will identify ways a computer is used in the home, at school, and in business.
STIMULUS ATTRIBUTES		
A. Format:		An incomplete statement or a question.
B. Content		Requires student to identify a computer application at home, school, or business.
C. Directions:		1. Select a way the computer is used (at home, school, or business); or 2. Computers can be used at (home, school, business) to ...; or 3. In (homes, schools, business) computers can be used to ...; or 4. Which does NOT need a computer? *or* 5. A computer is NOT needed to . . . .
RESPONSE ATTRIBUTES		
A. Format:		Short phrases.
B. Options:		1. Correct Response: Applications that can (or cannot) be done by a computer at (home, school, business).

From *Test Item Specifications for Minimum Student Performance Standards in Computer Literacy: Grades 3, 5, 8, and 11.* Tallahassee: Florida Department of Education, 1987. Reprinted by permission.

Table 3.8 *(continued)*

	2. Other Options: If question asks for an application that can be done by a computer, other options will be applications that cannot be done by a computer; and vice versa.
SAMPLE ITEM	Computers can be used at school to ● A. help students practice math. (1) B. erase the blackboard, (2) C. sharpen pencils. (2) D. clean the windows. (2)

the wording of the skill to that of each item type, one will recognize that each of these types contains information specific to a particular portion of the skill. Additionally, notice that the sample item specification contains stimulus characteristics and response characteristics for each content type. Study the detail in this table of item specifications, attending especially to the differences between it and the examples of test content specifications presented earlier in Tables 3.3, 3.4, and 3.5.

The first detail to notice when comparing test content specifications with item specifications is that the item specifications are typically much more detailed and prescriptive. While it was noted earlier that sometimes test content specifications identify the format to be used for an item, they do not prescribe characteristics beyond this. In contrast, item specifications are extensively prescriptive, including dictating characteristics of the stem and features for the response alternatives. Notice in the item specification presented in Table 3.7, for example, that the items are constrained by the specification to have exactly four response alternatives, and stimulus characteristics and response options are described. Further, each item type is accompanied by a sample item.

Another example of test item specifications is displayed in Table 3.8. There are obvious similarities and differences between the sample item specifications given in Tables 3.7 and 3.8. One similarity of special importance is the degree to which the content of the test is specified and the stimulus and response attributes are described. One difference between the item specifications is the format, content, and directions statements. Of course, each test development situation will need slightly different item specifications.

The Role Of Test Item Specifications

Test item specifications are usually intended as a working document internal to a test's development. They are not designed for examinees or test users as a guide to a test's content, nor are they pertinent to instruction. Lest there be confusion on this point, it is important to state that providing information to examinees and others about the particular content and features of a specific test is appropriate. In fact, such information may be necessary to fulfill the purpose of some domain-referenced and criterion-referenced tests; however, such a guide to the content and features of a specific test is not the same as a statement of test item specifications. Item specifications are intended primarily for the item writers so that their work will be focused in a way that reflects the test developer's intentions.

Item specifications are especially useful in tests that require a large number of items which are constructed by several item writers. In these instances, they add a consistency of approach to the task, ensuring that no one writer will prepare items that are incompatible with those prepared by the others.

A final point needs to be made about item specifications: One test item specification describesjust one item. Test item specifications are not usually meant to describe the features of several items varying in format, content, process category, and other features.

Happily, it is not necessary to write as many item specifications as there are items planned for a test. The same item specification may be used several times in a test. In fact, test developers commonly prepare only the number of item specifications equal to the number of cells on a two-way grid of test content specifications. This is all that is usually needed because in each cell for a table of test content specifications, the test developer stipulates that a certain content aimed at a specified level of cognitive processing may be assessed by several items. This makes it necessary to prepare only one item specification which can apply to all items within that cell. The idea of reusing an item specification several times in the same test should markedly reduce the number of separate item specifications needed.

Caution Against Making Test Content and Item Specifications Too Prescriptive

In recent years, specifications for writing test items (especially items intended for use in domain-referenced tests) have sometimes been prepared in too much detail. Some item specifications have attempted to describe nearly every conceivable delimitation to an item. Enormous amounts of time and energy are spent in preparing elaborately detailed item specifications. Even W. J. Popham (personal communication, April 1987), an early and vigorous advocate of thorough test content and item specifications, has retreated from using laboriously detailed specifications. Test item specifications are helpful only to the extent that they convey useful information to item writers. Beyond conveying needed information, they become hindrances rather than helpful aids.

Very narrow or laboriously defined specifications can lead to trivial items. In fact, an item writer's creativity may be inhibited if the limits imposed by the item specifications are too constraining. Frequently, brief item specifications are all that is necessary. A delicate balance needs to be struck between providing item writers with explicit, carefully considered ingredients and limits in writing test items and overly restricting a writer's ingenuity and creativity. A simple analogy may be made to using a road map to guide one in driving to a given destination as contrasted to a route ticket on a bus which dictates the exact roads and time of departure and arrival, leaving no room for deviation. Test item specifications should be more like a road map that can lead one toward a given goal without dictating every aspect of the course to be taken. When test item specifications are carefully prepared and skillfully used, they should aid the writer's task of making an item consistent with its intended objective.

But, simultaneously, another, ominous trend regarding test content specification is emerging. This occurs in some modern tests that seek to employ formats other than multiple-choice items, such in many performance assessments and "authentic" tests. Too often, developers of these tests eschew content description at all, relying instead only on gauzy terms, like "problem solving." Tests of this type also need well-thought-out and have carefully-articulated test content specifications. This point is expanded in Chapter 6, on constructed-response and performance assessments.

MAKING AN ITEM CONSISTENT WITH ITS SPECIFICATION

The Importance of Item-Objective Congruence

Achieving the maximum degree of congruence between the knowledge, skill, or ability actually assessed by an item and the intention for that item (as articulated by its specification) is of paramount concern to the item writer. It directly affects the interpretation of an item for content-related evidence of validity, and it influences the error of measurement, or reliability. Hence, the item-objective match is highly important. Unfortunately, determining the strength of that match is difficult. It involves an awareness of the full implications of the objective or skill to be assessed as well as a nearly complete command of the subtleties of the language selected for the item. Several examples will illustrate this point. The first example will be obvious; subsequent examples will convey various subtleties of the item-objective congruence problem.

To begin, suppose that Illustrative Item 3.1 is designed to gauge a student's skill is solving simple problems that involve proportions. As can be seen from a casual inspection, this problem involves recognizing units of measure and is unrelated to proportions. Hence, the item-objective congruence is lacking. This item is inappropriate for assessing this objective because of this mismatch.

Illustrative Item 3.1

Which is equal to 5 pounds?

A. 80 grams
B. 16 ounces
• C. 80 ounces
D. 7,000 grains

In contrast, consider Illustrative Item 3.2, which is meant to assess the same skill of using proportions. This item does require the examinee to use proportions, and thus exhibits a high degree of congruence to the objective. Of course, in this simple comparison of Illustrative Item 3.1 to 3.2, we are not concerned with whether the level of item difficulty is appropriate to the

Illustrative Item 3.2

Bill spends 45 minutes each day exercising, Karen exercises for 15 minutes a day, and Jack exercises 60 minutes daily. What is the ratio of how long Bill exercises to how long Jack exercises?

A. $\frac{45}{120}$

B. $\frac{45}{60}$

C. $\frac{60}{45}$

• D. $\frac{45}{15}$

examinees, as well as other considerations for good items; rather, it is used only to make the point of item-objective congruence.

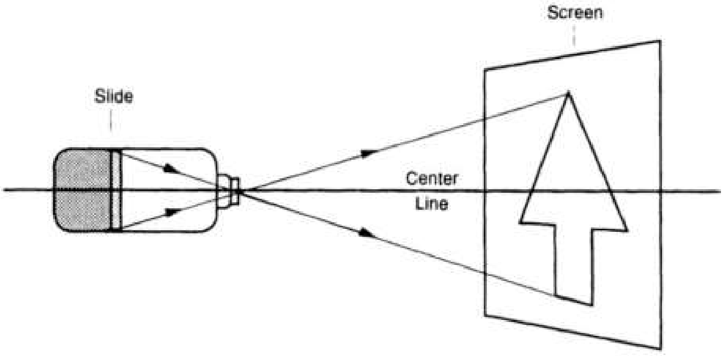
Another point warrants attention when looking at Illustrative Items 3.1 and 3.2. Although Illustrative Item 3.1 cannot be used for the specified purpose of assessing this objective because of its lack of congruence to the objective, the item may not need to be discarded out-of-hand. It may be matched to another more appropriate objective. In that new context, the item could suit its intended purpose.

Next, consider another example of item-to-skill congruence by again looking at two items designed to assess the same skill, Illustrative Items 3.3 and 3.4. The objective for the items is to classify the type of lenses used to produce particular images depicted in ray diagrams. One item is poorly matched to the skill while the other is well suited to it. Can you tell which is which? In this example, difficulties with subject content cause the mismatch for Illustrative Item 3.3 but not for Illustrative Item 3.4.

First, examine the item that is poorly matched to the objective, Illustrative Item 3.3. Despite the seemingly impressive graphic, this item does not assess the skill because it gives a ray diagram and asks about the image. The graphic is therefore inappropriate to the question asked in the item's stem. If the stem had asked about what lens was used, the item probably would be fine. However, the problem does not exist in Illustrative Item 3.4, which is accurately matched to its specification.

Illustrative Item 3.3

Use the diagram to answer the question below.

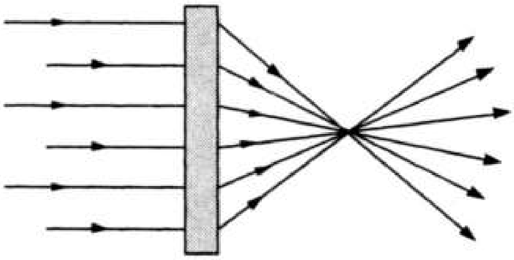


What type of image is shown in the diagram?

- • A. virtual
- B. fragmented
- C. real
- D. negative

Illustrative Item 3.4

Use the diagram to answer the question below.



Which lens would produce this light pattern?

- A. convex
- • B. concave
- C. colored
- D. polarized

It should be obvious to the reader from these examples, especially the preceding two illustrative items, that it takes a thorough knowledge of the subject matter to determine whether congruence exists. Often it is necessary to seek the advice of subject specialists to verify the content.

In addition to problems with content difficulty, the degree of congruence between a particular item and an objective is sometimes not obvious for reasons of language. Subtleties of language, such as shifts in word meaning, can be difficult to detect. For example, consider Illustrative Item 3.5.

Illustrative Item 3.5

Read the passage and answer the question that follows.

The soft, confident, quiet way in which Sam Carr spoke made Alfred start to button his coat nervously. He felt sure his face was white. Sam Carr usually said, "Good night," brusquely, without looking up. In this six months he had been working in the drugstore Alfred had never heard his employer speak softly like that. His heart began to beat so loud it was hard for him to get his breath.

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

• A. abruptly
B. quietly
C. nervously
D. shyly

The objective intended for the item is a commonly used one: Determine a particular word's meaning from the context of a given paragraph. Here, however, the word *brusquely,* which usually has one meaning (i.e., "abruptly"), by the context has another (i.e., "quietly"), thereby causing it to be incongruent with its intended skill. If, as intended, *brusquely* is an unfamiliar word at this level, the examinee will have trouble. The meaning is not evident from the context as the skill stipulates. In fact, the context would suggest response alternative B as the answer rather than A.

Practice in Determining Congruence

As can be seen, examining items for congruence often requires intense scrutiny of the item and its objective. By now, the reader probably realizes that judging a mismatch is not always obvious. Practice will help attune one to the complexities and nuances of language that affect an item's degree of congruence to an objective. To assist the reader, a few more examples are offered to help one recognize the subtleties involved in gauging item-objective congruence, as well as to garner an appreciation for this important ingredient of item validity.

Consider an objective that is typical of many in achievement tests: Solve problems requiring estimation in consumer, geometric, and physical situations. Illustrative Item 3.6 is meant to assess this objective. But does it really? Study this item and determine whether you think it does.

Illustrative Item 3.6

Which is the best estimate in feet of the circumference of a circular pool with a radius of 2.2 feet?

A. 5

• B. 15

C. 25

D. 50

At first glance, one may be tempted to believe that this item is a good match to the objective. After all (one may reason), the match exists because both the objective and the item contain the word *estimation.* However, upon close inspection of the item, one realizes that in order to arrive at a correct answer, examinees would not estimate. In fact, examinees are required to calculate. Further, the item assumes that examinees know the formula for circumference and the approximate value for ***pi* (*π*),** knowledge not specified in the objective. Rounding off ***π*** to 3 and using 2.2 for radius gives an answer of 12, which is not an option. Using 3.14 for *pi* and 2.2 for the radius gives an answer of 13.816, which is much closer to 15. It becomes clear upon scrutiny that this item is indeed confusing. The complexity of thinking and subtleties of language explained in these sample items are common.

Now, examine the problem of item-objective congruence by looking at three similar items, all intended to assess the same objective. We will see which item is best suited to this objective: Study the causes of soil erosion in an area, analyze the problems, and choose the appropriate action to prevent further erosion. Read Illustrative Item 3.7 and decide whether you believe it provides a good measure of the objective.

Illustrative Item 3.7

The Dust Bowl in the Great Plains region was the most dramatic example of wind erosion in the United States. How have people in that area prevented this from happening again?

A. planting trees to form windbreaks

B. plowing regularly with tractors

• C. irrigating the region

D. building many houses

Close comparison of the item with the objective shows that the item is a poor measure of the objective. The objective asks for an analysis of methods, but Illustrative Item 3.7 is a historical question rather than an analysis. The stem might just as well ask which action would best prevent wind erosion—the response alternative with "wind" (option A) being correct.

Now, consider the same objective but addressed by another item, Illustrative Item 3.8. This item has greater congruence to the skill than did Illustrative Item 3.7, but it still does not achieve very strong congruence. In this item the actual skill required to answer the item is more general than the one the item is supposed to assess. The stem might just as well have asked which action would increase erosion and left out the notion of its frequency (viz., "usually").

Finally, consider Illustrative Item 3.9. This item is a good measure of the objective because it is well-matched. It presents a situation particular to an area (as is stipulated in the objective), it requires an analysis (another stipulation), and it presents response alternatives which are plausible solutions to preventing further erosion. It is a good test item.

Illustrative Item 3.8

How do human activities usually increase soil erosion?

- • A. deplete soil fertility
- B. prevent natural erosion
- C. remove the vegetative cover
- D. increase the sediment in streams

Illustrative Item 3.9

In a generally flat, sandy area, which is the best method of conserving soil?

- A. clearing forests
- B. controlling weeds
- C. plowing grasslands
- • D. planting shelter belts

From these examples, one can see that determining the degree of congruence between a particular test item and a given objective or skill is neither obvious nor easy. It requires careful attention to the objective and to the item. A practiced eye will help. Such attention and practice are important requisites to becoming a writer of good test items.

CONCLUSION

This chapter presented information necessary for determining the content for test items as well as discussion of why this information is important. It described the basic tenets of validity and explained the relationship of these concepts to writing test items. As discussed, understanding these issues is an integral part of determining the content for test items.

Also, determining the content for items implies a scheme for organizing the content. To that end, this chapter offered a thorough description of test content specifications and item specifications. Finally, and importantly, there was a rigorous discussion of the importance of congruence between an item and the objective which it is intended to assess. Several examples elucidating various aspects of this consideration were offered.

This chapter and the preceding one described characteristics of items with special attention to their theoretical underpinnings. Such a complete understanding of items sets the stage for the following chapters, which focus upon several important practical considerations for writing items and editorial rules for formatting items. When set in a theoretical context, practical steps have meaning and can be undertaken with confidence that they will yield items of merit. We are now ready to turn to such practical considerations.