5

Developing Tests for Evaluating Instruction

Constructing a test may seem burdensome. The prospective test maker certainly should consider many things, and preparation of a good test of any instructional aspect of music is not simply a matter of writing a few questions. Testing in some form nevertheless is essential for evaluating the results of musical instruction. This chapter discusses determining objectives and what to test, aspects of item writing and analysis, and certain considerations involving musical examples. It is oriented toward paper-and-pencil rather than performance measures, which are covered elsewhere.

OBJECTIVES

Definitions

An *objective* is a relatively specific target toward which an organization or an individual works. One may have a biological objective, such as obtaining a drink of water, a learning objective, such as becoming able to make an oboe reed, or an affective objective, such as choosing to hear a favorite piece of music. In all cases, there is a specific want or need toward which purposeful behavior is directed. A *goal* is a more long-range target, one that requires attainment of numerous objectives to reach. One may have a goal of living a long and healthy life, becoming an oboe

virtuoso, or obtaining daily pleasure. Little, if anything, is gained by arguing whether a particular desired behavior or set of behaviors is an objective or a goal; it is a matter of relative specificity of the target and the observer's perspective.

Behavioral Objectives

A useful outgrowth of expanding educational technology and curriculum revisions in the 1960s is the *behavioral* objective. Mager's (1962) landmark publication was followed by numerous other works; Boyle (1974) prepared a summary compilation regarding behavioral objectives applied to music education. Behavioral objectives certainly were no educational panacea, but neither did they dehumanize education. They *are* valuable for the test maker because they suggest ways of demonstrating particular knowledge or attitudes that lend themselves to test items.

The key to a behavioral objective is the verb. Objectives calling for students to "recognize," "appreciate," "know," or "understand" are nonbehavioral because they do not indicate how an observer will "know" that the students recognize, appreciate, know, or understand. Objectives calling for students to "write," "list," "sing," or "clap" are behavioral because an observer can observe the specified action. Behavioral objectives may be more useful if they include criterion statements, as in

> The student will clap correctly at sight at least 90 percent of the rhythm patterns, on a list compiled by the instructor at a tempo of ♩= 72.
>
> Given a random list of Italian terms denoting tempo, the student will rewrite them in order from the slowest to the fastest in no more than one minute, with no more than two errors.

The "clapping" and "rewriting" are the observable behaviors, of course, and the objectives specify criteria for the behaviors and conditions under which they are to occur.

Behavioral objectives are not limited to relatively straightforward cognitive and performance behaviors. Elusive affective and aesthetic objectives can be stated in behavioral terms; the objective writer must decide what he or she will accept as evidence that the desired affect or aesthetic experience occurs. For example, what might a 13-year-old boy *do* if he "appreciated" a variety of musical styles? He might spend a portion of his allowance on recordings,

few of which are alike. He might attend concerts of varied performances, even when he is not excused from school to go. He might frequently retune his radio as he seeks changes in background music. In conversation, he might name composers and/or performers in each of several styles. All such activity can be observed. If particular examples of observable behaviors are not valid indicators of appreciation, the objective writer can select others—if "something happens" when appreciation occurs, that "something" can be manifested in some observable way. If no valid behavioral indicators of some desired condition exist, either the condition does not exist or it is useless as an instructional objective or goal.

SAMPLING BEHAVIORS AND OBJECTIVES

In any instructional situation beyond the most trivial, many events, relevant and irrelevant, occur. Many instructional outcomes are planned; some are spontaneous. It rarely (if ever) is possible to test everything, so what material should be tested? Of all the conceivable relevant behaviors, which ones should a measurement procedure sample? Instruction and evaluation thereof should be guided by comprehensive objectives, but decisions regarding the relative importance of objectives and potential test items inevitably are subjective judgments.

Deciding What's Important

What are the "really important" things to test? Of all the material pertinent to an instructional unit on, say, Beethoven, what did the instructor teach? In his or her best professional judgment, what does the instructor-testmaker believe the particular students should "know" about Beethoven and/or his music? Are biographical data, such as Beethoven's lifespan and approximate age at the onset of deafness important? Should students be prepared to recite a catalog of Beethoven's works by categories? Should they recognize excerpts from Beethoven's works, and, if so, how specifically should they identify the excerpts' sources (i.e., titles, movements, theme numbers)? All such decisions ultimately fall to the person responsible for instruction: He or she is obligated to prepare a test in such a way that it reflects the content as it was presented. Content validity is essential for tests that purport to

evaluate instruction; the student who studies Beethoven in accordance with the instructor's requirements should recognize the Beethoven of the test as the Beethoven of the classroom. Biographical and publication data and thematic recognition should be important during testing to the extent to which they were important during teaching.

Although this chapter is oriented primarily toward testing achievement that results from specific instruction, the question of what is important applies to aptitude, affective, and performance measures as well as classroom achievement measures. On the basis of knowledge, experience, and a sampling of available research, the would-be test maker must answer questions such as: What must a person do to be a musical success? How essential are nonmusical indicators of potential musical success? What are the indicators of a positive attitude toward a school music program? In what musical styles should a comprehensive musician be able to perform? These and similar rhetorical questions have no simple or general answer—the test maker must decide subjectively what to measure in the particular situation.

Note that from the standpoint of test construction "what is important" refers not to whether something should be *taught* or *characteristic of people,* but rather to what should be *tested, given* that something *has* been taught or *may* be characteristic of people.

Comprehensive Treatment of Important Aspects

Once he or she has decided what is important, the test maker is obligated to represent that importance in a comprehensive manner. All the objectives related to a particular corpus of material require some representation on a test of that material. A list of objectives, chapters in a text, or segments of a study guide must be represented in a balanced manner that fits constraints of time and the sophistication of the learners.

Constructing a detailed test plan can enable the test maker to cover the material sufficiently. Such a plan requires representing the corpus of the material to be tested, perhaps in the form of a list of objectives, numbers of pertinent pages in books, or a set of study questions. It also requires allocating the objectives, pages, or questions among particular test items or sections.

For example, suppose that an examination is supposed to cover three chapters of a music appreciation text. The chapters are respectively 30, 27, and 18 pages long. If the test maker determines

that the test should contain 36 multiple-choice items balanced among the chapters, a preliminary plan might reason that the total number of pages is 75, so the chapters should occupy respective proportions of .40, .36, and .24. Therefore, 40 percent of the test (14 items) should relate to the first chapter. Similarly, 36 percent (13 items) should relate to the second chapter, and 24 percent (9 items) should relate to the third. Within each chapter, one can construct a list of key points to determine the content of specific items. Adjustments in chapter proportions may be made if the test maker decides that there are more important points in a shorter than in a longer chapter.

As another example, consider a unit on basic properties of sound waves. Objectives may pertain to the students knowing (a) a definition of a wave, (b) factors influencing the frequency, speed, and direction of a wave, (c) components of a wave, and (d) conditions distinguishing a wave resulting from a musical tone from other waves. A test maker could itemize each major point and write corresponding objective test items. Possibly one comprehensive essay item requiring description of a wave's origin, its travel, and events that may occur as it travels would suffice. If multiple test items are necessary, proportional representation of perceived relative importance should determine how many items to write regarding each point. Naturally, there should not be slavish adherence to number—if the test maker believes one objective, such as indicating that a sound wave is the travel of disturbances rather than particles, is especially important, that objective may merit more test items than some other objectives.

This section's important point is that the test maker should *lay out the priorities.* The test must reflect what has been taught and how it was taught.

Taxonomies

A *taxonomy* is an ordered classified grouping within a subject field. In education there is the well-known Bloom (1956) taxonomy of cognitive objectives and the Krathwohl, Bloom, and Masia (1964) taxonomy of affective objectives. Psychomotor taxonomies are less well known; Colwell (1970) presented a largely perceptually based psychomotor taxonomy developed by Elizabeth Simpson at Illinois. Harrow (1972) developed a largely movement-based psychomotor taxonomy. All such taxonomies represent attempts to order possible educational objectives along a continuum ranging from simple

tasks requiring very little skill to complex tasks requiring extensive experience and use of intellectual, affective, or physical resources.

The Bloom taxonomy identifies six major categories of objectives: knowledge, comprehension, application, analysis, synthesis, and evaluation. There are subcategories within all but the application category, and even sub-subcategories within the knowledge subcategories. The simplest type of objective relates to knowledge of specific facts, such as knowing that an oboe is a double reed instrument. The most complex objective requires judgment in terms of external criteria, as in evaluating a set of published school orchestra programs to judge the balance of literature to which the students are exposed.

The Krathwohl et al. affective taxonomy runs from receiving through responding, valuing, and organization to characterization by a value or value complex, with subcategories in all categories except the last. Passive willingness to listen to "classical" music exemplifies the simplest level. Electing to listen to music when several other options are available indicates more than passive interest. A person who virtually organizes his or her life around opportunities to perform or listen to violin music would exemplify characterization by a value or value complex.

Simpson's psychomotor taxonomy begins with perception and is followed respectively by set (mental, physical, or emotional), guided response, mechanism, complex overt response, adaptation, and origination. The category boundaries are somewhat blurred; perceptual tasks can be rather sophisticated. Furthermore, as the term "psychomotor" implies, the cognitive aspect of a motor behavior may be extremely difficult to analyze.

Harrow (1972) developed a six-level taxonomic model based on a continuum of movements, discriminations, and manipulations. She defines psychomotor behaviors as "observable voluntary actions or action patterns performed by the learner and designated by the educator as being an essential portion of the educational goal of his [sic] particular curriculum" (p. 32). The first level, called Reflex Movements, contains respective levels of Segmental, Intersegmental, and Suprasegmental Reflexes. Harrow's second level, Basic-Fundamental Movements, includes Locomotor, Non-Locomotor, and Manipulative Movements. In order, Kinesthetic, Visual, Auditory, and Tactile Discrimination plus Coordinated Abilities comprise Perceptual Abilities, the third level. Physical Abilities—Endurance, Strength, Flexibility, Agility—constitute the fourth level; the fifth level, Skilled Movements, includes Simple, Compound, and Complex Adaptive Skills. Harrow's sixth and highest

taxonomic level includes Expressive and Interpretive Movement together as Non-Discursive Communication. There are subdivisions within most divisions at all levels, and even sub-subdivisions within a few subdivisions. As with all taxonomies, there are inevitable overlaps. Cognitive and affective aspects will relate to many voluntary movement-based behaviors.

One can waste considerable time in attempting to classify particular objectives or test behaviors exclusively and/or exhaustively into taxonomic categories; most musical behaviors involve complex combinations of domains and categories. Nevertheless, taxonomic structures are valuable for the test maker because they indicate various levels of knowledge, ability, and/or interest, all of which may suggest particular item types. Testing recall of many musical facts may indicate the degree of knowledge about those facts, but it may say little about how the test taker can *use* those facts. Testing interest in attending a particular concert for which the test taker may be excused from school will say little about a long-standing attitude toward music. Behaviors that fall relatively low in the taxonomic structure may be tested objectively and precisely; multiple-choice, completion, or observation items can easily be constructed to assess knowledge of musical facts or receptivity to a stimulus. Examining a student's evaluation of a musical composition or commitment to a value requires that the examiner make a more subjective judgment. As one considers various item types, one should be aware that the types of items and scopes of particular items may be suggested by a taxonomic structure.

ITEM WRITING

The advantages and disadvantages of various item types are discussed below, and the authors offer some general suggestions based on their experiences. The item writer must consider carefully the background and sophistication of the intended test takers as well as the test's purpose in relation to instruction. The needs for reliability, validity, and adequate coverage of objectives exist regardless of item type.

Wesman's (1971) chapter on item writing is a detailed discussion of short-answer, alternate-choice (true–false), correction, multiple-choice (in several versions), and matching items. The present authors' approach is focused toward assessment of musical behaviors, but it acknowledges the comprehensive treatment by Wesman, as well as Gronlund's (1981) principles for item construction.

Multiple Choice

The multiple-choice item is ubiquitous in educational and psychological measurement. It is highly useful for testing specific knowledge and application thereof. Although the emphasis here is on cognitive items, multiple-choice items may assess attitude or affect when a series of answers ordered on a continuum can be constructed. Multiple-choice items are easily scorable by machine; hand scoring is not especially formidable with group sizes less than 50, although obtaining desired test statistics may be.

Multiple-choice items consist of an opening question or stem followed by a set of alternate answers. Although items may be written to have more than one correct answer, if the examinee is told that he or she is to indicate all "correct" answers, it is customary to have one correct or "best" answer.

There is little or no agreement regarding the optimal number of choices. In theory a greater number of choices should reduce the likelihood of guessing the correct answer and increase the reliability. There are practical limitations regarding test and item length; extra choices (often called "distracters" or "foils" as well as "wrong answers") require extra reading time, and writing additional plausible choices can become difficult. Wesman (1971, p. 101) indicates that four- or five-choice items are the rule, except for younger (approximately six- or seven-year old) children. Ebel (1969) demonstrated that the greatest increase in reliability comes from an increase from two to three choices and that the longer the test, the less the effect of increasing item choices. Asmus (1978) demonstrated that reducing the number of answer choices would not necessarily reduce reliability because more items could be written. The present authors believe that the fewer-choices, more-items option is desirable. Three choices generally are quite sufficient for most multiple-choice tests.

Given the appropriate content area, one must consider certain caveats in writing multiple-choice items. Consider the following item:

> The nineteenth century saw the development of an extensive instrumental form known as the symphonic poem; one of that form's leading composers was
>
> A. Richard Strauss.
> B. Johannes Brahms.
> C. Felix Mendelssohn.

The brief music history lesson is unnecessary. It adds reading time, and in the event that any choice was not active in the nineteenth century it might provide clues to test-wise students. If the test writer wants the examinee to select Strauss from among a group of nineteenth century composers on the basis of the symphonic poem, all he or she should write as a stem is "A leading composer of the symphonic poem was". If the test writer also wants to assess knowledge that the symphonic poem arose in the nineteenth century or that it is an instrumental form, those concerns may be addressed in separate items.

Extraneous words within item choices lengthen test forms and are inefficient. Consider:

> The two main factors determining the reverberation time of a room are
>
> A. the absorption in the room and the shape of the room.
> B. the absorption in the room and the volume of the room.
> C. the shape of the room and the volume of the room.

It is much more efficient to say:

> The two main factors determining a room's reverberation time are the room's
>
> A. absorption and shape.
> B. absorption and volume.
> C. shape and volume.

The principle is to avoid excessive articles and prepositional phrases and state any information that applies to all answers in the item stem.

Grammatical and linguistic cues may lead a test-wise respondent to the correct answer. Incorrect use of an article can hasten elimination of an otherwise carefully planned foil, as in

> Japanese music often is associated with playing a
>
> A. oboe.
> B. aulos.
> C. koto.

Answers A and B are grammatically incorrect; the simplest solutions are to substitute *the* for *a* or incorporate the correct article in each answer. A bit more subtle might be the item

The composer of *Rigoletto* is

A. Georges Bizet.
B. Richard Wagner.
C. Giuseppe Verdi.

A student sensitive to the appearance of Italian names might make the educated guess that the only Italian composer available is the correct answer. Providing Italian foils, such as Rossini and Puccini, would solve the problem and might require more musical discrimination.

It generally is stock wisdom, although not clearly authenticated through research, that the following points should be observed in constructing multiple-choice items:

1. Individual answer choices that appear substantially longer or shorter than the others should be avoided.
2. A predictable answer pattern (e.g., ABC, ABC, ABC) should be avoided.
3. All available letters should be used more or less equally; for example, one should not write a test of 40 ABC items in which C is used only 5 times.
4. Answers such as "all of the above," "none of the above," and "some of the above" should be avoided unless there is a specific reason for encompassing or eliminating the designated answers; they should not be used to avoid writing additional substantive foils.
5. Obviously absurd choices, in relation to the test takers' knowledge and sophistication, should be avoided because they effectively narrow the number of choices.*

The authors strongly recommend instructing the examinee to select the "best" rather than the "correct" answer, especially on tests where there is legitimate debate about the ultimate truth of an answer. Furthermore, this allows for asking learners to recognize synonyms or alternate ways of stating particular facts.

A related argument-avoiding recommendation is to introduce an item with "According to . . ." when one specific but not universally accepted theory or viewpoint is the subject of the item. To ask

The number of choices in a multiple-choice item should be
A. 3.
B. 4.
C. 5.

*If providing a touch of humor overrides total concern for psychometric excellence, a test writer may include a few comical answers in a test intended for his or her own students.

is unfair because all three choices are defensible (see earlier discussion). If a particular authority took a position, it would be acceptable to ask "According to . . . , the number of choices . . . ." "According to" items are particularly useful in reference to a designated text, to what a particular music critic has said, or to what any particular authority (including the test writer) believes.

A well-constructed multiple-choice test can require substantial knowledge of specific information and the ability to use it. While one's communication skills may not be enhanced by selecting rather than constructing a response, the skillful limitation of selections to plausible but not quite "right" answers in addition to "right" answers is a time-honored way of assessing acquisition of information. In introductory music appreciation classes, in general music situations, in music theory, indeed in any situation where relatively specific recognition and use of musical terminology is necessary, the multiple-choice item is valuable. One sacrifices creativity, expression, and synthesis for precision and relative objectivity.

Matching

Relating one subcategory to another within one general category is the basis for matching items. Particular examples of matching items on music tests might include matching composers with compositions, measurement units (e.g., Hertz, decibels, cents) with tonal properties, performers with performance media, and compositions with styles. Here is an example of a matching item deliberately constructed to exemplify the genre positively:

> Match each musician in the left-hand column with the musical medium or style for which the musician is best known, according to our class discussion and listening assignments, from the right-hand column. Only one letter should be placed in a blank; it is possible that the same letter would fit more than one blank correctly.

____	1. Johann Strauss	A.	opera
____	2. Merle Haggard	B.	symphony
____	3. Johannes Brahms	C.	jazz
____	4. Richard Rodgers	D.	musical comedy
____	5. John Philip Sousa	E.	country-western
____	6. David Brubeck	F.	rock
____	7. Robert Schumann	G.	cantata
____	8. Giuseppe Verdi	H.	musique concrete
____	9. Michael Jackson	I.	marches
____	10. J.S. Bach	J.	waltzes

In the above item it is clear that only one letter is to be placed in a blank. While it is possible to require more than one answer (if Mozart were listed, it would be difficult to justify forcing a choice between A and B), the examinee should know if that is a possibility in the particular matching item. Although the number of blanks and choices is equivalent (it need not be), there is no way that one answer may be right automatically by elimination (B should be used twice; H fits none of the listed musicians). The examinee is aware that simple elimination will not work because he or she is told that a letter could be used more than once. The length is manageable (one should avoid having a matching item requiring more than one page), and the distance from each blank to the correct choice varies. The item is relatively homogeneous in the context of a comprehensive survey of musical styles and musicians, including composers and performers from different eras.

Linguistic cues can be a problem in a matching item, as in matching an author and a music history source or a composer and a composition on the basis of foreign language similarities.

Matching items are a useful way to investigate whether learners can make a correct series of relationships or associations within some logical category. They are as objective as multiple-choice items and can easily be scored by hand, although some respondents' printed or cursive letters may be ambiguous. Machine scoring may be difficult without adaptation of score sheets.

True-False

True–false items may be deceptively simple to write. Lehman (1968, p. 31) discusses the difficulties of constructing nonambiguous true–false items and the relative ease with which a person might guess the correct answer. Wesman (1971, p. 92) also discusses the ambiguity problem and recommends (p. 94) "unless there are truly compelling reasons to do otherwise, the test constructor would do well to favor other forms of items." Despite authoritative warnings the true–false item remains popular in some teacher-made tests, so it is discussed here.

A good true–false item must be defensibly true *or* false in the context in which it is used. The truth or falsehood of "Sibelius was a great musician" could be assessed in the context of his *not* being a carpenter, playwright, or general—then it is obviously true. However, one could focus on "great"—"great" in whose opinion? Sibelius is revered in Finland and certainly is well known to some

orchestra enthusiasts, but was he "great"? His "greatness" was and is too ambiguous to judge on a yes–no basis. "The clarinet is a band instrument" is, of course, true in the sense that concert bands usually contain clarinets, but the clarinet also is an orchestral, a chamber music, and a solo instrument. "The saxophone is an orchestral instrument" is false, unless one's orchestral experience includes works such as Ravel's orchestration of Mussorgsky's *Pictures at an Exhibition,* Ravel's *Bolero,* Milhaud's *La Creation du Monde,* or Prokofiev's *Lieutenant Kijé Suite.* Qualifying words can decrease ambiguity, as in

> Sibelius's career was that of a musician.
> Some conductors, performers, record producers, and other musical authorities believe that Sibelius was a great musician.
> Concert bands are very likely to contain [or not contain] clarinets.
> The saxophone usually is not [or is] part of a symphony orchestra.

Care should be taken that qualifying words do not make true items long in relation to false ones.

That which is not completely true is false. That point must be made to examinees, especially younger ones.

In the interest of reliability it may be worthwhile to print T–F in front of each item for examinees to circle. Then there is no question of whether a person has written or printed a T or an F. Some test makers may prefer that examinees write "true" or "false," but that procedure is slightly more time consuming. With machine scoring of answer sheets, true–false items may be treated as two-choice multiple-choice items.

Guessing *may* be discouraged by using a guessing formula that penalizes a wrong answer more than an omitted answer; a ? option may be provided for the unsure student. (However, see Chapter 3.)

True–false items may be combined with other item types. An essay item could be built around an explanation of why a particular statement, even an ambiguous one, is likely to be true or false. Examinees may be asked to rewrite false statements, thereby combining true–false with short answer items and requiring response construction as well as classification. One point may be awarded for each correct decision, and, for each false item, one point may be awarded for a correct rewrite. When this option is elected it is wise to underscore or italicize particular words for which change is allowed (in true as well as false items!) in order to avoid a situation such as

> "Allegro ma non troppo" means "fast but gradually slow down"

being rewritten to become true as

> "Allegro ma non troppo" does not mean "fast but gradually slow down."

Yet another option is to turn a true–false item into a type of multiple-choice item by requiring true–false decisions about several segments, as in

> The solo concerto for a wind or string instrument was a
> A
> common form in the eighteenth century; one of the principal
>
> composers of solo concerti was Franz Schubert, and another
> B C
> was Wolfgang Mozart.
>
> A. T F
> B. T F
> C. T F

In general, the authors believe that the difficulties of true–false items outweigh their usefulness. In a conventional multiple-choice item, the examinee really is evaluating the truth or falsehood of each particular choice, yet the "best answer" criterion allows selection of the "most truthful" (or most defensible) answer. While acquisition of musical factual information can be assessed via true–false items, confusion of facts with likely misunderstandings is better assessed through multiple-choice items. Construction of a response is better assessed through the item types we now discuss.

Completion

In a *completion* item, the test taker supplies a word or words to fill a blank or blanks in a sentence. Obviously, the given statement must contain enough words so that a knowledgeable person would be able to see what is required. All blanks should be of equivalent length, and there should be no grammatical cues.

> Consider the following completion items:
> 1. _____ composed *The Rite of Spring.*
> 2. The four major woodwind instruments in standard symphony orchestra instrumentation (i.e., "woodwinds in pairs") are the _____, _____, _____, and _____.

3. A solo concerto is most likely played by a solo instrument and an _____.
4. _____ was (is) a composer of atonal music.

All of these items "look" all right; the blanks are of equal length, and particular words would complete the sentences. Closer examination indicates that the first two items only could be answered correctly in the expected way, but what does one do with partial answers or misspellings? Is "Igor" alone sufficient for the first answer? The third item clearly contains a grammatical cue that "band," "quartet," "chorus," or "piano" is not the answer. Does one accept "ensemble" or "accompaniment"? The fourth item's use of "was (is)" forces the examinee to consider living as well as dead composers, but there clearly is more than one correct answer that the test scorer must be prepared to accept. What happens if the examinee is familiar with an obscure composer of atonal music who is unknown to the scorer?

Clearly, completion items can tap knowledge that is organized logically in a sentence format. While obviously ambiguous and overly leading [to the correct answer] items can be avoided, the test maker must be prepared for unexpected but correct answers and must have a consistent procedure for handling alternative spellings. The need to interpret answers makes using different readers to score identical tests less reliable. Machine scoring is impossible, unless the test is administered via computer-controlled word processing and text editing.

Completion items may be especially valuable in the context of formative evaluation. Completed answer blanks may be compared for alternative correct answers. If knowledge of exact spelling and ability to write terms are considered important, completion items are viable. However, they generally are impractical for testing large groups, and their use in summative evaluation where a grade or a curricular decision is at stake should be approached with considerable caution.

Short Answer

What the authors call a *short-answer* item is a request for a relatively small amount of specific information in an unstructured format. There are varying degrees of specificity, and the amount of information can vary. When a short-answer item requests a narrative description of specific facts, as in requesting a citation of five twentieth century composers of art music with a brief indication of

why each is important, the item may in fact be a short essay item. As with completion items, ambiguity must be avoided and the reader must be prepared to evaluate unexpected but veridical answers. Here are a few examples:

1. Name four composers active during the baroque era.
2. List in score order the four voice ranges commonly found in an adult mixed chorus.
3. List six songs that we sang in class this semester.

The second and third items have definite answers. The first also does, but an examinee may name an obscure (to the scorer) but nevertheless "active" baroque composer.

It is possible to have multisectional short answer items. An item could request four composers and three compositions by each. A student could list instruments or voices and notate a playing or singing range. (In a way, the time-honored exercises of melodic and harmonic dictation are short-answer items; if some notation is provided they may be completion items.)

A type of relatively structured short-answer item provides the information in a scrambled order. If one wants a student to "know" tempo markings, he or she could request a list of the terms (in which case it is only fair to specify how many terms) and definitions, test definitions for each term (or terms for each definition) through multiple choice or completion, request matching of terms with definitions, or provide the terms and request an ordering, as in

> Arrange the following terms in order from the slowest to the fastest:
>
> Andante
> Allegretto
> Largo
> Moderato
> Presto
> Allegro

Short-answer items can assess how well the examinee can produce something rather than only recognize it. They may be useful where the test takers should "know it cold." They are relatively quick to write, but they are time consuming to score. (Machine scoring is not possible except with sophisticated equipment.) The number of possible points may vary among short-answer items; the scorer must determine how many points are possible and on what basis each point will be awarded. When time permits and construction of a

detailed list or set of exemplars is deemed important, short-answer items are potentially useful.

Essay

There are reasons to avoid essay examinations as a means of evaluating instruction. Different readers may assign widely varying ratings to an identical essay. A fluent writer may write an elaborately impressive but shallow answer. A person who has learned a considerable amount may not be able to show it because he or she can not write well or the corpus tapped by the item does not require much of what was learned. An essay may be read with a positive or negative bias because of what the reader knows about the writer; this is the "halo effect." The essay item nevertheless is a traditional examination mode, especially in higher education, and a well-constructed essay can be a valuable learning experience. There are particular steps in constructing essay items that can enhance the quality of an essay examination.

The "halo effect" may be avoided by having essay writers identify themselves only by a code to which the reader will not refer until the essays are read and evaluated. Test takers could use student or social security numbers. They might designate a particular combination of letters, numerals, and symbols to which the reader will not refer until the essays are scored. An assistant instructor or paraprofessional might remove names and assign each paper a code number that is not revealed to the evaluator until the scoring is complete. Having all students write on identical paper with same-color ink may prevent some identity clues. In small classes, handwriting recognition may be an insurmountable problem unless typing is feasible.

It is important that the essay writer have a fair chance to know what is expected and to organize the essay into an acceptable form. This requires a certain amount of structure in the "question." Consider the following examples:

1. Discuss how a person hears and organizes music.
2. Indicate how to tune an instrument.

While a student might know more from the context of the examination and previous instruction than is evident here, these are both "wide open" items that could elicit divergent answers. Is the "discussion" of a physiological or perceptual nature? Is one to "discuss"

the basic hearing process? Is one to consider different sources of music? Does organization mean conceptually or in some physiological way? Does organization mean conception of musical structures, such as melody, classification of styles, recognition of form, or something else? *What* instrument is to be tuned, or does it matter? Is "have someone tune it for you" an acceptable answer? Are tuning mechanics all that is required or is one supposed to evaluate different types or frequencies of tuning standards? The items are far more clear regarding what is expected when written in this way:

1. (Part A) Describe the basic hearing process involved in listening to music. Account for reception and transmission of the sound wave through the ear, perception of tonal properties, and relaying a sensory message to the brain.
   (Part B) Describe several principles by which a listener organizes auditory input to the brain as music. Consider melody, harmony, form, and rhythm.
2. Describe the procedures by which you would tune a brass instrument. Account for a tuning standard, mechanical adjustments to the instrument, and adjustments made by the player while playing.

While room for diversity remains, what is expected is far more clear as a result of the item writer's guidelines.

A model answer constructed by the test writer may guide the reading of the essays as well as reveal unexpected sources of difficulty or a need to provide further guidance. Model answers also can provide useful feedback to the students.

Essay examinations should be read while the reader is reasonably alert and free from distractions, which requires allowing adequate time. One should read question by question rather than examination by examination and avoid assigning final grades until all responses to a particular essay item are read. Evaluative criteria should be based on accuracy, organization, and completeness of the requested information, not on length or neatness. If time permits, a rereading of the essays and/or reading by more than one knowledgeable person to arrive at a consensus may enhance reliability.

When a group of readers are to score an essay independently, guidelines may be useful. Sample "good," "bad," and "indifferent" answers may be provided; a list of events or phenomena that the writer should discuss may focus the readers' attention. In general, all responses to a particular essay item should be read by all read-

ers to avoid introducing further extraneous variance or bias, as would be the case if only the "hard" graders read one item while a mixture of "hard" and "easy" graders read another. If there are too many papers for any one person to read all responses to a given essay item, some other testing procedure should be employed.

Unless the essay must contain an exact amount of specific facts, scoring an essay item is a matter of judgment and/or categorization. Scoring usually is based on matching an overall impression of the essay with a score or grade rather than on counting right answers. An ABCDF or ESU (excellent–satisfactory–unsatisfactory) classification could be employed. An essay could be allocated a maximum number of possible points, with that number being an upper limit, as in scoring a gymnastics event or round of a prize fight. Generally grades or points should not be assigned until all the essays are read unless criteria are detailed and rigid. Much of the material in Chapter 8 is relevant to making a "judgment call" on an essay.

Verbal skills, written as well as spoken, are crucial in higher education and many careers. There comes a time in the study of music theory, history, acoustics, psychology, sociology, and pedagogy when it is important for the student to synthesize, evaluate, and prioritize in a way that cannot be done "objectively." The essay examination thus is essential, despite its difficulties. Coffman (1971, p. 300), in his treatise on essay examinations and their inherent problems, concludes:

> There is little doubt that for many purposes, particularly when large numbers of candidates are to be examined and when the measurement involves subject-matter knowledge, objective testing is more efficient than essay testing. On the other hand, when questions are carefully constructed and focussed on performances involving the communication of complex relationships among ideas and when carefully planned scoring procedures designed to minimize systematic and variable scoring error are applied, essay examinations can provide reliable and valid evidence of achievement.

WHAT ITEM TYPES SHOULD BE USED?

The types of items that a test writer employs are a function of the subject matter, the learners' experience, developmental stage, and sophistication, and the test writer's objectives and beliefs. Multiple-choice items are especially useful in many situations and may be

the best overall item type for assessing factual information and specific relations among facts. General musical knowledge, fundamentals of notation, facts of music history and theory, pedagogical questions, and anything where logical accurate and inaccurate responses to stimuli can be categorized are appropriate areas for multiple-choice items. Where construction of a response is desirable, as it may be when accurate use of musical vocabulary and musicians' names is of concern, completion and short-answer items are appropriate. If synthesis and application of information from diverse sources is important, there is no substitute for an essay examination. It is possible to mix item types within one examination; a music history final examination might include multiple-choice items assessing relations of factual information, matching items regarding composers and compositions, and an essay item regarding stylistic contrasts and similarities among three compositions. In any case, the test writer is responsible for professional judgments regarding item selection and construction, and he or she must be prepared to defend those judgments, with consideration of the learners and the subject matter.