1995

5. Systematic Item Writing And Test Construction

Anthony LaDuca
National Board of Medical Examiners, tladuca@nbme.org

Steven M. Downing
American Board of Internal Medicine, sdowning@uic.edu

Thomas R. Henzel
American Board of Internal Medicine

SYSTEMATIC ITEM WRITING AND TEST CONSTRUCTION

Anthony LaDuca

National Board of Medical Examiners

Steven M. Downing

American Board of Internal Medicine

Thomas R. Henzel

National Board of Medical Examiners

Standardized objective testing remains the most popular mode of licensure testing. Even where other types of tests are incorporated, it is often the case that they are provided as complimentary to standardized, multiple-choice (MC) tests. Moreover, scoring theories and standard-setting procedures have been developed over the years in the context of standardized MC testing. At the same time, critics have pointed to limitations of contemporary MC testing practices, including lack of fidelity to real-life challenges and emphasis on recall of factual minutiae. In our view, testing professionals should make conscientious attempts to modify test development procedures so as to address valid criticisms. In this chapter we offer several suggestions for improving licensure test development, although it may not be feasible to adopt the entire array of recommendations we make. We are providing an intentionally wide selection in the hope that testing professionals will find something of use in their field of practice. Our discussion emphasizes careful design and systematic item-writing methods. We describe types of test items and make suggestions for development and maintenance of an item pool. Later we discuss test-construction procedures.

OPERATIONAL ASSUMPTIONS

We assume that the testing program is intended for use in licensing persons who are entering an occupation or profession in a U.S. jurisdiction. Our

discussion assumes further that the program is new; however, the implications for already established licensure programs may be clear to the reader. The testing programs we consider are those that rely on paper-and-pencil techniques generally associated with standardized testing. These imply having examinees fill in spaces on answer sheets that are optically scanned at a later time. We are also assuming that the standards for passing the licensure test will be established using one or more of the content-based approaches that are presently available. Such standards are fixed and maintained through equating proceduresusing the appropriate statistical methods. Details of these procedures are provided elsewhere in this volume. In this chapter we assume that systematic pretesting of newly written multiple-choice questions (MCQs) will be implemented as part of the testing program.

Much of our experience has been in the context of licensing and certifying physicians and our examples are largely restricted to medical applications. We believe that the features we outline will be effective with nonmedical professions as well.

IMPORTANCE OF TEST DESIGN

Test development comprises the full array of activities associated with bringing a standardized assessment into operation. The particulars of what we designate as *design* are of special significance in development of licensure tests for two reasons. First, the imperative to assemble evidence in support of the content validity of the examination is heightened in the licensure context. Second, the logical and procedural linkages between the design and the test items must withstand close scrutiny.

Job Analysis, Job Relevance and Content Validity

Content validity retains a somewhat controversial character among measurement specialists. Much contemporary commentary relegates content validity to an inferior status because it is described as emerging from the apparent fit between the test content and the persons (i.e., experts) involved in the development of the test. This version of content validity places it outside the preferred paradigm of interpretations of examinee scores. In our view this disparagement of content validity is unwarranted in licensure testing. Validation of licensure tests may rely heavily on evidence of unimpeachable "job relevance" of test content, but there is no reason to exclude empirical processes from content validation, including interpretations of scores. More to the point, the imperative to establish the unimpeachable job relevance of the licensure test enhances the importance of design because it is at the level of test design that the issue of relevance is first addressed.

The job relevance perspective implies that the test items in the licensure examination must be linked through systematic means to a well-defined representation of the demands of the occupation or profession. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1985) call for a "job analysis" in licensure test development (Fine, 1986)

and this has come to be a well-accepted element of the process (see chapter 4). Although we prefer an alternative method to conventional job analyses, the more significant point is the imperative to start with a representation of the target occupation or profession. The purpose of such a representation is to establish a definition of knowledge and skill that is essential to competent practice. It is possession of the candidate's knowledge and skill that the licensing examination is intended to establish or confirm, and the presumption is that the public is protected by such an assessment.

Among the available alternatives for job analysis, we prefer representing the target profession by devising a model of the situations that comprise the professional domain. This strategy has evolved from a social constructionist view of professions, which argues that the knowledge and skill possessed by competent practitioners is displayed in response to the demands posed by encounters in a real-world (i.e., social) environment (LaDuca, 1980; LaDuca, 1994; LaDuca & Engel, 1994). Therefore, an effective means of laying out the knowledge and skill demands of an occupation or profession begins best by defining the situations that constitute the domain of the occupation or profession.

This approach is responsive to the special context of physician licensure, wherein there is tension between the increasing specialization of physicians during their extended training, on the one hand, and the language of licensure laws, which usually emphasizes the credentialling of *undifferentiated* practitioners, on the other hand. Our response to this dilemma has been to devise a method for representing the generalist practitioner, although such persons are largely hypothetical. For other professions this dilemma may not exist. Nevertheless, we are impressed that the approach we have devised over the years retains significant advantages for other professions as well.

Our approach involves constructing a *practice model* based principally on log-diary surveys of practitioners in which they report their activities. It is important to note that the practice model captures crucial elements of professional situations in order to describe them. There is no attempt to presume modalities of intervention in the professional situations. In professions where alternative interventions are available, (e.g., psychotherapy), the practice model approach only asserts the imperative that qualified practitioners engage successfully with, for example, married couples considering a divorce, or treatment of a child displaying school phobia. Different and acceptable modes of treatment are defined in the subsequent analysis of the allowed situations.

Decisions about the content of the licensure test are made by a committee of recognized experts in the field, but in this approach their decision making is informed by the structure of the description of the practitioner's work as derived from empirical data. The design of the licensure test then results from the informed judgments of content experts who have evaluated the data underlying the practice model.

For example, surveys of selected office-based physicians, supplemented by other data bases, lead to a practice model that identifies the character of the patient population and the nature of clinical problems encountered. These data have shown

that a large majority of physicians' office-based clinical encounters are with patients who have been diagnosed previously and who are presenting in the context of continued care. In the face of these data, content experts have agreed that the test blueprint should incorporate a continued care frame in a majority of test items. At the same time, the expert committee has not endorsed a simple one-to-one correspondence between the blueprint and the specific clinical problems and diseases reported in the surveys, because that would imply a physician licensure test focused on patients seen for general physical examinations and upper respiratory infections (i.e., "colds"). There may be instances where rarely occurring, but high-impact problems may be preferred over frequently occurring, low-impact conditions. Thus, the practice model approach retains reliance on expert judgment about the weighting of content on the licensure examination. The logic of that process puts the experts in the position of interpreting data descriptive of the professional domain and devising rationales for appropriate departures from the weightings implied by the empirical data. (For a more complete treatment of the manner in which this process leads to test specifications, see LaDuca, Taylor, & Hill, 1984.)

The composite of expert decisions, informed by an empirically derived practice model, establishes the main points of the content of the licensure test, although the benefits of these analyses would be diminished if the writing of test items was not carried out in a systematic manner. In the following sections we describe several approaches to systematic item writing. In the section on "Developing the Initial Item Pool" and in the appendices we illustrate the ways in which the job analysis, evaluative objectives, and test items are connected. We begin by identifying types of objective items used in licensure and certification examinations. Examples of these item types are provided and their strengths and limitations described. In the interest of completeness, constructed response items also are discussed.

SELECTED RESPONSE ITEMS

Objectively scored selected response items are the most frequently used item type on standardized licensure and certification examinations. Selected response items require examinees to choose an answer from possible answers supplied as a list of options. This family of item types has been in use for at least the past 50 years and, at its introduction, virtually replaced the constructed response item.

There are several types of selected response items currently in use: *single-best-answer* questions, *true-false* questions, *matching* questions, and *extended-matching* questions. Single-best-answer items require examinees to choose the one best answer from among a list of options or possible answers supplied by the test writer. The various matching formats are variations of the single-best-answer format. The most popular item type in use today is the multiple-choice question (MCQ) with four or five options and one option keyed as correct (Type A). The alternate-choice (AC) item, a special case of the MCQ, presents a stem question with only two possible answers (Downing, 1992; Ebel & Frisbie, 1986). The strength of the AC item is that it can test content that does not require absolute truth or falsity, such that the more correct option is selected. Matching and extended-matching items are also used in large-scale examinations.

Current practice is to designate ("key") only one option as correct in high-stakes examinations using selected response items, although it is possible to create good test items that involve more than a single correct response. In some contexts these may be preferable, as when equally attractive treatment options may exist for selected illnesses, or several appropriate diagnostic studies should be pursued. Classical test theory is most efficient for single-best answer items (e.g., Ebel & Frisbie, 1991); it is less well suited to items with more than one keyed response. The literature shows efforts to develop scoring methods that accommodate items with more than one correct response, principally item-response theory and polychotomous response models (e.g., Embretson, 1984). Testing professionals also must be sensitive to validity problems that may arise because of examinees' lack of familiarity with this response format.

True-false questions require examinees to respond to the truth or falsity of statements or questions. The stand-alone true-false item is rarely used in standardized examinations, but multiple true-false (MTF) items are employed. MTF items present a statement or open-ended question in the stem and require examinees to respond "true" or "false" to each of the varying number of options presented. Each true-false item in the set is generally scored as right or wrong, although some testing programs use various "cluster" scoring procedures for these items.

In the next section these selected response formats are discussed in turn, with an example of each item type given, and the format's strengths and limitations noted.

Multiple-Choice Questions

Where multiple-choice questions are used for licensure and certification examinations, the single-best-answer MCQ is the format of choice. The MCQ format presents a question or incomplete statement in the item stem and several (typically four or five) options as possible answers; only one option is keyed as the correct answer.

The most useful test for following the activity of disease in a patient with rheumatoid arthritis is

(A) erythrocyte sedimentation rate
(B) serum antinuclear antibody titer
(C) serum protein electrophoresis
(D) serum rheumatoid factor concentration
(E) synovial fluid antiglobulin titer

Strengths

Multiple-choice items permit efficient and straightforward measurement of cognitive knowledge and educational achievement. Because responses are easily machine scored, large-scale testing can usually be accomplished in a cost-effective manner. Although MCQ testing has been criticized for emphasis on simple recall and trivia, it is possible to measure complex knowledge, such as judgment, decision making, and synthesis of knowledge (Maatsch, Huang, Downing, & Munger, 1984). MCQs are time-efficient for both the item writers and test developers, and also for examinees challenged by these items. The research base and psychometric theory for MCQs is very rich.

Principles of MCQ construction are discussed widely (e.g., Haladyna, 1994; Haladyna & Downing, 1989a; LaDuca, Staples, Templeton, & Holzman, 1986; Roid & Haladyna, 1982). However, the empirical research on aspects of these item-writing principles is somewhat less rich. (See Haladyna & Downing, 1989b, for a good summary.)

MCQ Weaknesses

MCQs require examinees to recognize and select correct answers that are supplied. Presentation of answers may clue the correct answer, making this task less difficult than constructing responses to questions. Some research supports this belief (e.g., Ebel, 1972), but recognizing correct answers and constructing correct answers are very highly correlated. Nevertheless, implications for validity of using MCQ testing for licensing continue to receive constant scrutiny.

All selected response formats allow the possibility of the examinee guessing the keyed correct answer when the correct answer is unknown. In general, providing a larger number of options lowers the probability of randomly guessing the correct answer. Because of the possibility of guessing, MCQs traditionally have four or five options.

In our view, psychometric concerns about guessing are excessive. If guessing were a large source of error variance for MCQs, reliability estimates would be much lower than typically reported for such examinations. When sufficient numbers of items are used, the guessing issue becomes trivial. Licensure and certification examinations should use large numbers of test items for content validity and high reliability. Lord (1944) reported that the three-option format is the optimum for high-ability examinees. Lord (1977) replicated these findings using item-response theory. Haladyna and Downing (1993) report that even well-written four- or five-option MCQs used in national certification and standardized college admissions examinations have only two distractors that perform as expected, effectively creating a three-option MCQ.

Other potential weaknesses of MCQs include ambiguity, bias, reading level problems, security problems, testwiseness clues, and test anxiety. Ambiguity is reduced by careful and thorough editing by both content experts and professional test editors. Various techniques to identify and reduce test bias are available (Cole & Moss, 1989). Reading level must be appropriate to the examinee population and is controlled by careful editorial review and pretesting. Test security is problematic for MCQs; to ensure the valid interpretation of test scores, MCQ examination materials must be secured throughout the test development process including test administration and scoring.

Much heat and little light have been generated by issues of testwiseness, coaching and its effects, and test anxiety issues. Examinees must be familiar with MCQ formats. The *Standards* (AERA, APA, & NCME, 1985) require that examinees have the opportunity to practice with item formats prior to the certification and licensure examination. Coaching probably has some small effect, (see Chapter 3, Rosenfeld et al.) but far less effect than thorough study of the content measured by the examination and a smaller effect than the statistical effect of

regression toward the mean (e.g., Becker, 1990; Smith, 1991). Test anxiety may affect test scores for some examinees, but this phenomenon, if it exists, is not limited to the selected response formats.

Matching Items

Matching questions present several test items that are answered by selecting from a set of (usually) four or five options. Matching sets may be very useful for testing examinees' knowledge of related concepts and conditions. In contrast to single-best-choice items, matching items should have options that are of apparently equal likelihood. In the medical context, selecting the most likely diagnosis is a good example. (It is possible to use more than two stems for each matching set.)

The most likely explanation is:

(A) Conversion disorder
(B) Dysmorphic body image
(C) Malingering
(D) Normal behavior
(E) Panic disorder

1. A 66-year-old woman comes to the clinic requesting evaluation for breast cancer after a close friend and neighbor was diagnosed with the disease. Mammography is arranged. Later, the patient is relieved when results of her mammogram are negative.

2. A 21-year-old woman comes to the clinic. She says that she was on the way to an acting audition when "I got a racing heart, I couldn't breathe, I got dizzy and I was afraid I was going to die!" She says that this type of episode has happened three times before but never this bad.

Matching Item Strengths

For the most part, matching items share the strengths noted for MCQs. Traditional matching items may be most efficient for testing comparisons and relational concepts across broad topic areas.

Matching Item Weaknesses

Recall of facts and their relationships may also be the limitation of traditional matching items. The focus is narrowed by the theme (e.g., diagnosis) and the items must pose classic presentations if examinees are to make the distinctions. It also is difficult to write matching items that measure higher-order knowledge because of the possibility of word associations cuing the examinee to the correct response. The comparison of concepts usually requires that their distinctions be less subtle; it may be imperative to limit the contrasts to black-and-white distinctions.

Extended-Matching Items

Matching items are a variation of the single-best-answer question format. In the traditional matching item, questions are to be answered by selecting from a lettered list of possible answers. A newer variation is the extended-matching item (Case, Swanson, & Stillman, 1988; Case & Swanson, 1993). Extended-matching items have four essential components: a common theme, a lead-in, a list of options, and two or more item stems.

COUGH

The most likely diagnosis is:

(A) Acute bronchitis
(B) Atelectasis
(C) Bronchial asthma
(D) Bronchiectasis
(E) Cancer of the lung
(F) Chronic obstructive pulmonary disease
(G) Cystic fibrosis
(H) Pneumococcal pneumonia
(I) Pulmonary embolus
(J) Pulmonary tuberculosis

1. An afebrile patient complains of "tightness or pressure" in the chest. He has dyspnea, a cough and expiratory wheezing.

2. During the past 5 years, a patient who smokes two packs of cigarettes a day has developed progressive dyspnea accompanied by coughing and wheezing.

Extended-Matching Strengths

The extended-matching format encourages item stems that provide more detail (e.g., in medicine, stems that present extensive clinical descriptions of patients) and provide for a longer list of options. The research data (e.g., Case & Swanson, 1989) suggest that this item format is more difficult than MCQs, with higher item discriminations, and higher reliability estimates; however, these findings probably are not universal. The item format lends itself best to diagnostic questioning, and therefore, probably assesses "higher" cognitive levels than the traditional matching format. Item authors seem able to produce large numbers of extended-matching items efficiently (Case & Swanson, 1993) and the format lends itself to the item-modeling principles outlined in this chapter.

Extended-Matching Weaknesses

In general, the limitations of matching items may be amplified when a larger number of options are used. Because a common theme is needed for the format, it is possible to oversample in some content areas while overlooking other content areas. Such over- and undersampling could reduce the content validity of the examination. Also, attempts by item writers to capitalize on the longer options list may lead them to develop questions that make trivial distinctions. Longer lists may allow for subsets to function as distractors for different questions, permitting more capable examinees to reduce the functionality of the entire array.

Multiple True-False Items

The multiple true-false (MTF) item presents a statement or open-ended question, followed by two or more related true-false items. The examinee is instructed to respond to each option as true or false. (This item type is sometimes referred to as the Type-X item.) Frisbie (1992) presents a comprehensive review of this item type and a summary of the research reports on this item type. An example follows.

The table shown below represents the performance of Test A for Disease X in 100 patients.

TEST A	DISEASE X	
	Present	Absent
Positive	50	8
Negative	12	30

Correct statements include:

(A) The sensitivity of the test is 81%
(B) The specificity of the test is 79%
(C) The positive predictive value of the test is 86%
(D) The negative predictive value of the test is 28%
(E) The prevalence of Disease X in this population is 58%

Multiple True-False Item Strengths

MTF items are consistently more reliable than single-best response MCQs, when reliabilities are adjusted for equal amounts of testing time (Frisbie, 1992). MTF items have been shown to be more difficult than MCQs in some studies (e.g., Albanese, Kent, & Whitney, 1977; Kreiter & Frisbie, 1989). Concurrent validity evidence (correlations of MCQ and MTF item data) shows that the two formats measure about the same knowledge (e.g., Frisbie & Sweeney, 1982). Criterion-related validity evidence for the MTF item is sparse. Albanese, Kent, and Whitney (1977) found that MTF items predicted GPA as well as other formats, such as MCQs.

MTF items are time-efficient for both examinees and item authors. Although there are exceptions, most timing studies (Frisbie, 1992) suggest that the ratio of MTF items to MCQs answered per minute of testing time ranges from about 2.3 to 3.4. Hence, MTF items are very efficient.

Multiple True-False Weaknesses

Downing, Grosso, and Norcini (1994) showed that, compared with MTF items, MCQ items had higher criterion-related validity for an independent external rating of competence. The MTF format typically lends itself to assessment of facts and other so-called "lower" cognitive taxonomic levels. For example, Baranowski, Downing, Grosso, Poniatowski, and Norcini (1994) show that in subspecialty certifying examinations in Internal Medicine, 40% to 80% of MTF items are classified as measuring knowledge, rather than judgment or synthesis.

CONSTRUCTED RESPONSE ITEMS

Constructed response items require the examinee to supply an answer rather than select an answer from a listing of possible answers. Constructed response items are currently used in some large-scale testing programs, such as the Medical College Admissions Test and the College Board's Advanced Placement Program.

Examples of constructed response items range from the familiar "fill in the blanks" items and short- and long-answer essay tests to complex computer-scored natural language items and computer administered and scored problem-solving exercises (Martinez & Bennett, 1992). Another example of a constructed response item is math problems that require the examinee to grid the computed answer on a special optical-scan sheet, which can be computer scored. Bennett (1991) offers a taxonomy of constructed response items ranging from the simple to the very complex.

Constructed Response Strengths

The principal strength of the constructed response item format is that examinees must supply answers rather than identify answers from a list. It is widely

thought that supplying answers is a more complex task than recognizing answers. The research evidence for this advantage of constructed response items is sparse, but constructed response is believed to require different skills than selected response formats (Bennett, 1991).

The constructed response item format eliminates clueing of answers, because the examinee must formulate an original response. This formulating of a response is believed to be a more complex cognitive task than merely recognizing the correct answer from a list of possible answers. Constructed response items also appear to pose more authentic real-life problem-solving assessments, because real-life problems rarely come with a ready-made set of possible answers. Also, constructed response items are often easier to construct than selected response items because there is no need to devise plausible distractors.

Constructed Response Weaknesses

Constructed response items are difficult to score reliably. Development of machine-scoring methods for these items is only in its infancy (Martinez & Bennett, 1992). In order to score paper-and-pencil constructed response items reliably it is generally necessary to use multiple raters or scorers and then average their ratings. Interrater agreement is the essential reproducibility required in this context. Raters must be trained and "calibrated" to their task and their performance must be tracked over time. Sample answers, that make explicit the range of correct and incorrect answers, must be developed. Obviously, the rating process itself is expensive and time-consuming. Expert judgment is often required, in which case raters may need to be skilled professionals in the content area, which may be even more expensive and logistically complex.

Much development is currently taking place in constructed response formats, including work in the higher technology areas of computer scoring of these items. For example, Martinez and Bennett (1992) describe a natural language computer-scoring system being developed by Kaplan (1992). In this system, constructed response short answers are scored by a pattern-matching computer program; high agreement is reported for the computer scoring and human judges. Another example of development in this area is the computer-administered "figural response" items used in architecture examinations. Martinez (1993) reported that the figural response item performed well, but was less reliable than parallel MCQs.

Another area of development using currently available technology is the so-called uncued item format (Veloski, Rabinowitz, & Robeson, 1993). Although not strictly a constructed response item, the uncued format uses multiple choice stems as questions, but the answers are selected from a very long list (1,000 or more options) of possible answers that are available for all items. Answer codes are then gridded on a special optical-scan answer sheet for machine scanning. This format may be considered a hybrid between selected and constructed response items, utilizing the strengths of both while minimizing the limitations.

Using Item Sets

The matching formats usually call for several items associated with a list of some sort. However, sets of items may also be used effectively in non-matching

formats. This tactic allows assessment of several aspects of the same general topic. A familiar example is the reading comprehension test, which presents a paragraph for the examinee to read, followed by several related questions that challenge the examinee to interpret what was read. This general format has been described by Haladyna (1992) as "context-dependent item sets," although there are other names.

Item sets are helpful in promoting assessment of higher-order thinking, because a richer problem or situation can be presented and several aspects tested. For example, in medical licensure testing, Dillon, Henzel, Klass, LaDuca, and Peskin (1993) have reported on their experience with the *case cluster*. This format consists of a series of four to nine single-best-answer MCQs related to a specific patient encounter. (See Appendix 3.) This format permits advancing the narrative of the encounter and posing challenges that reflect multiple aspects of the case such as initiating therapy, modifying therapy, making referrals to other clinical specialists, admitting the patient to the hospital, monitoring for progressive deterioration, detecting new problems in an established patient, and exploring ethical aspects of managing patients and their families.

DEVELOPING THE INITIAL ITEM POOL

The following section describes approaches to writing MCQs for assessing the knowledge of practitioners and students. The origins of this work reside in development of MCQs for tests used in evaluating the clinical knowledge of physicians and, for the most part, the examples cited are medical. The approach recommended here is believed to be equally appropriate for use with testing programs for other professionals.

Although the history of MCQs in standardized testing extends back more than 50 years, it has been only during the past two decades that systematic methods for writing MCQs have been advocated vigorously (e.g., Haladyna, 1991, 1994; Haladyna & Downing, 1989a, 1989b; Popham, 1978). Collectively, these methods have been described as an item- writing "technology" (Roid & Haladyna, 1982) that is intended to assist in production of larger numbers of higher quality MCQs. We will describe two methods that rely on making linguistic linkages between items and objectives. Separately, we will describe another method that permits development of large numbers of items based on exemplary items.

OBJECTIVES-BASED METHODS

All objectives-based item-writing methods start with a statement pertaining to an important aspect of knowledge or skill. These statements are assumed to have emerged from the job analysis procedure selected to support the design of the licensure examination. Our approach relies on content analysis of scenarios describing details of professional situations located in the practice model. We prefer this to soliciting knowledge and skill statements from expert practitioners in the target profession, but what follows is applicable to such descriptive statements as well.

In some applications, an objective is recommended for each item. However, this strategy may lead to an overabundance of objectives without commensurate gain in numbers of items or in quality of measurement. It is better to think of an

evaluative objective as broad enough to encompass a set of at least 10 related items. Such objectives may be thought of as domain descriptions. In this context, item writing becomes part of domain-referenced test construction (Baker, 1974). What is crucial to effective objectives-based item writing is making explicit connections between the language of the objective and the words comprising the item.

Preparing Objectives

Objectives-based item writing requires the identification of the content reference, or topics, eligible for inclusion. In the two approaches described here, the content reference is a separate listing, such as clinical problems or diseases. Strictly speaking, development of evaluative objectives (or domain descriptions) is separate from the process of objectives-based item writing. In fact, developing objectives probably should involve a different group of experts, though there may be overlap.

An effective method of preparing evaluative objectives has been used in selected examinations developed by the National Board of Medical Examiners (NBME). The method begins with a practice model or other framework for situations that the competent target practitioner is expected to encounter (Burg, Lloyd, & Templeton, 1982; LaDuca, Taylor, & Hill, 1984). These situations may be described in a brief scenario, written and reviewed by content experts. Content analysis of the scenarios identifies important objectives. In our test development work, the objectives have been related to a physician task (e.g., performing a physical exam; using diagnostic aids; managing therapy). The items written to assess these objectives generally require a clinical vignette that describes a specific patient. Because the goal of the physician licensure testing is to evaluate the examinee's readiness to practice medicine, this focus on patient management seems warranted. Other evaluation contexts may require alternative perspectives, but whatever the context of evaluation, the advantages of developing relatively few objectives with broad content boundaries remain. Examples of items written in an objectives-based manner are found in Appendix 1.

THE LEAD-IN METHOD

The *lead-in* is the name given to the sentence or phrase that ends the item stem. Functionally, the lead-in puts the question to the examinee. Therefore, the lead-in serves as the direct link between the evaluative objective and the test item. A lead-in may be in the form of a question (*"What is the most likely diagnosis?"*), or it may be in sentence-completion form. For example, if the objective relates to knowledge of appropriate diagnostic tests, then one reasonable lead-in might state, *"The most appropriate diagnostic study is. . ."*

It is recommended that one or more lead-ins be prepared when objectives are developed. With experience, additional lead-ins may emerge and these should be made available. Writing test items using evaluative objectives and lead-ins should proceed as follows:

1. Identify a clinical problem AND a related objective.
2. Select a specific lead-in that is associated with the assigned objective. If available, sample items should be provided as additional aids to effective item writing.

3. Confirm that the item's lead-in poses the question that relates to the referenced evaluative objective.
4. Write an appropriate stem preceding the lead-in addressing the selected clinical problem and including sufficient clinical detail (e.g., patient age, history, complaints, history).
5. Write the correct answer and distractors that are logically and grammatically consistent with the lead-in.

Appendix 2 contains a brief selection of evaluative objectives associated with physician tasks. In addition, one or more lead-ins are provided as examples.

THE AMPLIFIED OBJECTIVE METHOD

The amplified objective (Baker, 1974) is the most systematic method described here. It is also the most demanding. Amplifying objectives works best where objectives are plentiful and large pools of items are needed. It is effective when groups are responsible for instruction or evaluation, because the process emphasizes clear explication of content relationships. An amplified objective has four parts. They are:

1) General Evaluative Objective;
2) Sample Item—illustrates the results of the amplifying process;
3) Content Limits—identifies appropriate content by defining key terms in the objective;
4) Response Limits—describes item formats and testing conditions; states criteria for correct and incorrect responses.

The following section describes a modified process for amplifying evaluative objectives. Assessing cognitive aspects of clinical competence is emphasized, and so, in general, the items are clinical vignettes.

Amplifying Evaluative Objectives

1. Identify the focal *Evaluative Objective*. Use wording that states (a) what information will be provided to the examinee, (b) what action the examinee will take, and (c) what information the examinee will be acting upon. For example, the objective should have this structure:

> *Assesses severity of patient condition and makes judgment as to current status, prognosis, or need for further action. (Response options are inferences or conclusions referenced to the patient in complete sentences.)*

2. Prepare a *Sample Item.* Write or select at least one very good example of an item conforming to the amplified objective. Identify the keyed (correct) response.

3. Develop *Content Limits.* Begin by highlighting specific terms in the objective that identify, or imply, important clinical content. In the objective cited above, these would include *patient, acute but limited problem, ambulatory setting,* and *likely diagnoses.*

4. Establish *Response Limits.* Specify item formats (e.g., A-type, four-option). Also, elements of stem content should be delimited (e.g., patient age, presenting complaint, signs and symptoms, setting, etc.), and variations on lead-ins should be specified.

5. Define the correct responses, usually by referring to a content reference, such as a list of eligible diseases, drugs, laboratory studies, etc. Also, you should stipulate the character of incorrect responses. For example, if the correct response is a respiratory infection, you must decide if all distractors must be respiratory infections. You may insist that distractors be varieties of pneumonias, or that other etiologies may be represented.

THE ITEM MODELING METHOD

Pioneered at the NBME, this method is helpful when the goal is rapid expansion of a small item pool. The process begins with a high quality MCQ that can serve as a model for many similar items. The assumption is that a well-written item, relating to a complex content topic or domain, is only one instance of a larger "family" of equivalent items (Haladyna, 1994; LaDuca, Templeton, Holzman, & Staples, 1986; Shea, Poniatowski, Day, Langdon, LaDuca, & Norcini, 1992). Other members of the "family" can be developed by imitating, or modeling, the *source* item. To guide the modeling process, a set of specifications for new items is based on a content analysis of the source item. Item modeling produces large numbers of items, but in a limited content area. Item modeling is more successful with MCQs that have longer stems, especially clinical vignettes. Modeling basic science items has been less successful.

Item Modeling Process: Preparing Modeling Specifications

1. Select a *source item.* It should be a well-written MCQ, preferably a clinical vignette, on a topic for which you want additional items. Use a single-best choice (A-type) with 4 or 5 options as the source item.

2. Highlight the specific terms in the stem that are important clinical content, (e.g., clinical setting; patient age, sex, and race; medical history; presenting complaint(s); signs and symptoms; and results of diagnostic studies).

3. Identify the correct (keyed) response, and the *content category* to which it belongs. For example, the answer to the question may be a diagnosis; a follow-up diagnostic study; a decision to admit the patient to the hospital; a referral; a modification in the patient's medications; etc.

4. *Review the available wrong options (distractors),* and discard any that are inconsistent or flawed. List additional plausible alternatives, and, if possible, stipulate rules for combining choices in new items. These "distractor rules" should guide item writers by delimiting options that should, or should not, appear together.

5. *For each clinically important term in the stem,* list several significant alternatives. The alternatives should be "differences that make a difference" in the clinical context. For example, how would the clinical situation be different

- if the patient were a young child instead of an adult?
- if the patient were a woman instead of a man?
- if the patient had significant family history of disease?
- if the diagnostic studies produced different results?
- if the patient's prior treatments were different?

6. *Prepare complete specifications for each new item.* Identify the content of the new stem by labeling one clinically reasonable combination of the alternatives. Then, for each new stem, identify or provide a keyed response. Finally, for each keyed response, specify the desired distractor rule. Figure 1 shows a sample specifications table for a modelling procedure.

TEST CONSTRUCTION

In describing this systematic test development process, we have assumed that the examination is new and intended for a high-stakes decision; that the test specifications have been developed through a defensible and systematic design process; that content experts will develop the test items and create all the test materials; that a committee structure is in place to create and approve examination policy and plans, to review and approve content specifications, to write and/or review test items, and so on; that items will be pretested for all future forms of this examination; and, that items will be stored in an item pool to access for future examinations. (We must omit from this discussion the critical issue of content validation of test items, although it has great significance for checks of adequacy of test items as measures of important knowledge and skill. This topic, so crucial in licensure testing, is addressed more fully in chapter 4.)

It should be noted that test security is needed from the very outset of test development for high-stakes examinations such as those examinations used for licensure and certification. Procedures for securing the examination items while they are being developed and reviewed should be as thorough as those security measures used during and after examination administration. Secure mail should be used to move items from author to test agency to reviewers; computer systems must be as secure as possible and access to items must be limited to those with a need to know. The security plan for the examination should be developed together with and as an integral part of the test development plan.

Appointing Expert Panels

Because individual test items are the building blocks for examinations, a primary task is to select and train item authors. Several defensible models are possible, but selecting item writers who are expert in the content to be measured and who are invested in the success of the testing program are key elements. Item authors must be willing to follow item writing guidelines established for the testing program and make a reasonable effort to accommodate the timelines established for test development and review. The lead-in method and the item modeling techniques discussed in this chapter provide highly efficient means of generating large quantities of high-quality test items. Item authors can be readily trained in these techniques and typically find these methods useful. Generally, about one-half to two-thirds of the items written will ultimately survive all content and editorial reviews and pretesting.

Item-Writer Training

Item writers for the testing program must be thoroughly familiar with the guidelines for item development and all the procedures established for submission and review of items. Test security requirements for authors should also be well

STEM ATTRIBUTES				OPTION ATTRIBUTES		
PATIENT	**SYMPTOMS**	**PHYS EXAM**	**STUDIES**	**OPTIONS**	**KEY**	**DISTRACTORS**
0. 10-yr-old boy previously healthy	10-day progressive cough, low-grade fever, dyspnea on exertion	diffuse rales bilaterally	Chest x-ray (perihilar infiltrate)	A) Pneumonia due to respiratory syncytial virus B) Pneumonia due to *streptococcus pneumoniae* (Pneumococcus) C) Pneumonia due to *mycoplasma pneumoniae* D) Pneumonia due to staphlococcus E) Congestive heart failure F) Tuberculosis	C	1. Any four others. 2. Include all pneumonias 3. Include only two other pneumonias
1. Same as 0 above.	10-day progressive cough, 24 spiking fever and dydpnea on exertion	diffuse rales bilaterally; egophony on right	Chest x-ray (two small air fluid levels on right)	Same as 0 above	D	Follow rule 2
2. Same as 0 above	Same as 1	fine crackling rales in right posterior base with impaired resonance	Chest x-ray (infiltrate in right lower lobe, fluid in right fissure); WBC = 38,000; 96% PMN	Same as 0	B	Follow rule 3
3. 2-mo-old boy; normal delivery	10-day persistent cough, afebrile, alert, rapid breathing	red eyes with purulent discharge from both; diffuse rales bilaterally	Chest x-ray (hyperinfiltration, diffuse interstitial infiltrates	G) Group B beta-hemolytic Streptococcus H) Hemophilus influenzae I) Pseudomonas J) Chlamydia pneumonia K) Pertussis	J	4. Include cited options
4. Same as 3 above	10-day mild cough with choking episodes, low-grade fever, profuse mucoid nasal discharge	conjunctival injection; normal chest sounds	Chest x-ray (perihilar infiltrate, scattered atelectasis); WBC = 30,000; 70% lymphocytes	Same as 3 above	K	Follow rule 4

Figure 1. Sample Item Modeling Specifications.
(Adapted from LaDuca, Templeton, Holzman, & Staples [1986] Item modelling procedure for constructing content-equivalent choice questions. Medical Education, 20, 53-56.)

understood (for example, authors may not keep copies of their items; secure mail should be used to ship questions; FAX and electronic mail transmissions are not secure).

Generally, specific item-writing assignments to individual authors are helpful. Such assignments will specify the number of items to be produced, the type of item format, and the exact content domains in which items are to be produced (from the test specifications). Sometimes it is helpful to tailor the assignment to the specific content expertise and interest of the authors. Authors could reasonably be asked to produce 25 to 50 MCQs over a period of several months.

Typically, item authors can be trained to the item-production task in about a one-half to full-day workshop, during which time clear written instruction is given, with many good and bad examples of the item types to be used presented. New authors also should have the opportunity to actually write items, receive feedback on their attempts, and receive some practice in review and critique of other authors' items.

Item Production

Timelines of sufficient length should be established to allow adequate time for item writing, review, rewriting, editing, and approval cycles. Generally, a minimum time of 18 months is needed to initiate a new high-stakes testing program (from the start of the test development process to the first testing date).

Each item should be subjected to a systematic development process that includes initial development, review, revision, and pretesting (Hambleton, 1980). One such sequence is shown in Figure 2. According to this sequence, the item is produced by the author, following the guidelines and content assignments established. The assigned items are received by the test development agency, generally logged in, and then entered into a computer system (ideally tied to an item-banking system). Subsequently, newly written items are edited by skilled professional test editors who are familiar with test construction technology. All items should be reviewed for potential bias and insensitivity to population subgroups.

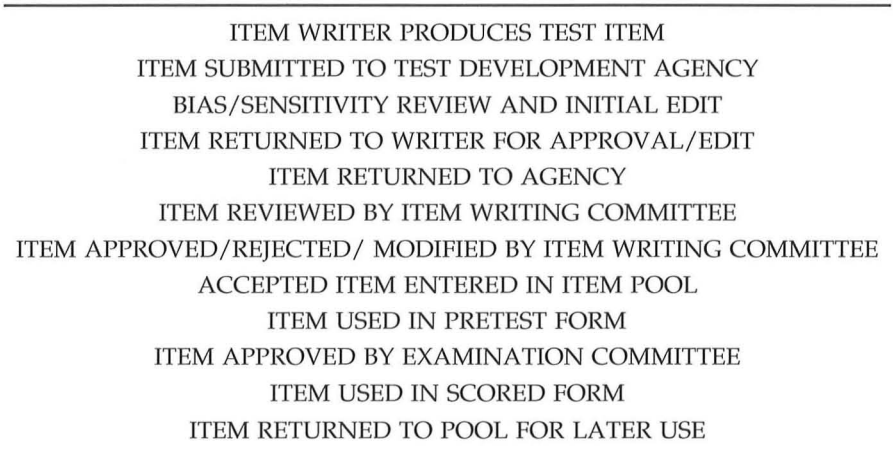

Figure 2. Life cycle of a test item

Edited items are then returned to authors for comment, clarification of questions raised by editors, and final author approval. Such items are then returned to the test development agency and prepared for content review. Content reviewers must be expert in the discipline and willing to review test items critically. It is preferable that reviewers have had experience as item writers because it will increase their sensitivity to the task confronting the item writers. Batches of test items can be securely mailed to content reviewers for critique and/or all items produced can be reviewed by a content committee charged with examination development. Reviewers, just like item authors, must be familiar with test security procedures and willing to follow all explicit security guidelines.

Item Pool

Once an item is accepted by a test development committee, the item is entered into the item pool and awaits pretesting. It is helpful to have rated the items for priority in pretesting. All identifying information about the items is entered with the item to facilitate test construction. An item pool can range in complexity from a simple paper system on which items and identifying information are stored on index cards to sophisticated, tailor-made computer software designed for an individual application. Many commercially produced software systems are currently available. Essential features of an item pool include: easy item storage and retrieval; the capability to store, sort, and retrieve items based on all relevant variables such as content classification, author, item statistics, and so on; integration with word processing and/or editing systems; and the flexibility to be modified easily as requirements change. (For more details about item banking, see Chapter 8).

Test Construction

Test construction refers to the actual process of building test forms from the item pool of approved items. For this discussion, we assume that we are building a new high-stakes examination to be administered in one day of testing time. The examination will contain a total of 200 MCQs for scoring and an additional 160 items for pretesting only. The examination is to be administered to 1,000 examinees. Four test booklets containing the same 100 scorable items, but including 20 unique pretest items, will be produced for the 4-hour morning session. The pattern will be repeated for the 4-hour afternoon testing session. Figure 3 illustrates this design.

This test booklet design allows 2 minutes of testing time per MCQ and permits a sufficient number of examinees (e.g., 250) to take each pretest item. For programs using traditional item and test statistics, about 100 examinees is minimum for each pretested item. For programs using IRT methods, the number of examinees may need to be much higher. Test booklets will be "spiraled" so that they will be distributed to examinees in the sequence Form 1, 2, 3, 4, 1, 2, 3, 4....n.

The purpose of pretesting is to generate score performance data on test items—to try out the item with examinees who are similar to those examinees who ultimately will be challenged by the item for "credit." Pretesting allows the test developers to select items that have the most desirable psychometric characteristics,

thereby enhancing test validity and reliability. It is important to restrict the number of pretested, unscored items seen by each examinee, but about 10% is a reasonable target.

Examination Administration, Scoring, and Evaluation

Once the examination is administered, answer sheets and all test materials are returned to the test development agency (using secure shipping methods) for scanning and scoring. Test materials are first checked in and any missing materials are traced and located. Answer sheets are machine scanned to produce an electronic file of the responses recorded by the examinee on the answer sheet. Scoring is accomplished by applying the approved scoring key to the response. (It is assumed that scoring programs are available and that all psychometric issues such as passing score determination, scaling, score reporting, choice of psychometric model, and so on, have been made prior to examination administration.)

A preliminary scoring and item analysis takes place, using carefully constructed and approved answer keys. A process of "key validation" may be completed prior to the final examination scoring. Key validation refers to a final verification of the scoring keys' accuracy by a group of content experts. (When all items have been previously pretested the key already has been validated; under these circumstances "key confirmation" may be a better name for this procedure.) This final key review is facilitated by reference to the preliminary item analysis data for each item. Criteria for item statistics such as item difficulty and discrimination are used to "flag" items for content review and key accuracy. For example, items that are very difficult and/or that do not discriminate well between those who score

TIME	TEST FORM	COMMON SCORED ITEMS	UNIQUE PRETEST ITEMS
Morning			
	A1	100	20
	A2	100	20
	A3	100	20
	A4	100	20
Afternoon			
	P1	100	20
	P2	100	20
	P3	100	20
	P4	100	20

Figure 3. Test booklet design for accommodating item pretesting.

highest on the test and those who score lowest may be flagged for evaluation. Content experts may decide to delete (score as correct for everyone) the item, change the key, or score the item as it was administered.

After final scoring, pretested items are evaluated. Item analysis data are examined for each pretested item using some predetermined criteria of item difficulty and discrimination. If the item meets the criteria, it is retained in the item pool for possible use on a scorable form of the examination in the future. (Items will be reviewed by content experts prior to use on a scorable examination.) Items that fail the statistical criteria for inclusion in the item pool may be discarded or returned to item authors or test development committees for evaluation and possible rewriting.

The performance of examination items is useful feedback to item authors. Some systematic method of item tracking should be included in the specifications of the item pool, such that the performance of items can be summarized for individual authors. Simple statistics such as the average difficulty of an author's items and the proportion of items passing the pretest criteria may be useful to authors as feedback.

Items used on a scored portion of an examination have some shelf life for possible reuse if the test remains secure. Shelf life depends on several variables: how rapidly the content and/or the test specifications change and evolve and how restrained content committees are in editing or otherwise modifying "used" questions.

Item pooling for high-stakes examinations require maintaining good items for possible reuse because creating new test material is expensive and labor intensive. As a general rule no more than 50% of items might be reused together from a previously scored examination (pretest items are not included, because these are "new" items); however, reusing about one-third of items is preferable.

One very basic reason for reusing items on an examination is to allow for the statistical procedure known as "equating." Examination equating (discussed in detail in Chapter 11) refers to the process of adjusting test scores on a current version of an examination in order to maintain the identical interpretation of the passing score from administration to administration. Equating allows one to interpret test scores in exactly the same way from administration to administration; it is as though all examinees took the same examination. Hence, when examination scores are properly equated, the meaning of the passing score is the same from administration to administration. No matter how carefully examinations are constructed (even from pretested and used items) it is impossible to maintain the identical average difficulty of the test from administration to administration. Equating solves this problem so that examinees are neither benefitted nor penalized by getting a slightly easier or more difficult examination.

A design of a typical classical measurement equating model used by many high-stakes examinations requires the use of a common set of used items (often referred to as "anchor" items). Because these common items are used to anchor the equating, such items must be unchanged from administration to administration. When the equating is carried out, the performance of examinees on these common

items is compared from the first to the second administration. This performance is used to adjust scores on the current administration of the examination to maintain the identical score scale.

Common items used for equating cannot be edited or changed in any way. Although there is always some creative tension between content experts and test development agencies around editing anchor items, the logic of equating requires that items be repeated in exactly the same presentation from administration to administration. Some effort should be made to retain as much common context as well (i.e., use in the same book). If used items are edited substantially (and this is where the debate often occurs), then such items should not be used as part of the equating link.

Conclusion of Testing Program

At the end of each testing cycle, it may be very useful to prepare a technical report of all relevant test development, administration, standard setting, scoring, and reporting activities. Such a report is an attractive method for maintaining records of activity in support of the program's defensibility. Summary psychometric analyses should be reported, including average item difficulty and discrimination, estimates of score and decision reproducibility, and mean scores and pass rates for important examinee subgroups. Specific recommendations and plans for improvement of the program should be included in the final technical report.

Program Audits

Madaus (1992) has advocated routine external review as a further guarantee that high-stakes testing programs are fulfilling their obligation to protect the public. The fundamental argument is that all testing programs can be improved by systematic and independent inspection by qualified professionals. The consequences to the public and to the profession may be too serious to restrict responsibility for quality assurance to persons who may have vested interests.

The auditors' primary responsibility is to the protection of the public. Therefore, it is imperative that auditors be independent of all interested parties and without any stake in the outcome of the audit. External, independent auditors should be highly qualified measurement professionals, with experience in the specific type of examinations being reviewed.

The *Standards* (AERA, APA, & NCME, 1985) provide the basis for all testing program audits. Schmeiser (1992) provides additional guidance concerning the ethical obligations of measurement professionals. The auditor should collect systematic data about all important aspects of the testing program; test development, item quality, item review and editing, content validity evidence, and test security should be examined. Additionally, it is important to evaluate psychometric data, including item analysis, statistical evidence of validity, estimates of reliability, procedures for determining passing scores, and score reporting. The evaluator should make specific recommendations for program improvements, with implementation of recommendations included in subsequent audits.

SUMMARY

We have covered substantial ground in this chapter. We have discussed several critical elements of test development for assessments used in licensing. We remain cognizant that the purpose of licensure is protection of the public and the profession from unqualified practitioners. Because these are high-stakes decisions, the developer is obliged to give priority to issues of quality, defensibility, and validity in all components of the testing program.

We have restricted our discussion to conventional methods of standardized testing, with emphasis on multiple-choice formats. Irrespective of the formats used, we have recommended systematic item-writing methods, relying on committees of content experts appointed especially for this purpose. We have assumed that the design of the program has been conducted in accord with current requirements as summarized in the *Standards*, with particular attention to the imperative to assess in areas of knowledge that are of unimpeachable relevance to the demands of professional practice. The content specifications for the examination must be delineated carefully and based on the implications arising from an appropriate job analysis. Detailed discussion of the methods for accomplishing this phase of the program development is beyond the scope of this chapter.

We have recommended that item pools be developed, consisting of large numbers of test items that have been pretested successfully. In addition we have urged the use of content-based standard-setting methods for establishing criteria for adequacy of performance. We have suggested maintaining fixed standards through application of statistical equating methods described elsewhere in this volume. Finally, we have admonished test developers and licensing agencies to exercise extreme caution in the maintenance of test security.

We began by acknowledging that critics of standardized testing make valid arguments in some instances. We believe that the overall quality of standardized licensure testing will be enhanced greatly by attention to the techniques and procedures detailed in this chapter.

REFERENCES

Albanese, M. A., Kent, T. H., & Whitney, D. R. (1977, November). *A comparison of the difficulty, reliability, and validity of complex multiple choice, multiple response and multiple true-false items.* Paper presented at the Annual Conference on Research in Medical Education, Washington, DC.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Baker, E. (1974). Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. *Educational Technology, 14*, 10-16.

Baranowski, R. A., Downing, S. M., Grosso, L. J., Poniatowski, P. A., & Norcini, J. J. (1994, April). *Item type and ability measured: The validity of multiple true-false items.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 9,* 179-190.

Bennett, R. E. (1991). *On the meanings of constructed response.* Princeton, NJ: Educational Testing Service.

Burg, F. D., Lloyd, J. S., & Templeton, B. (1982). Competence in medicine. *Medical Teacher, 4(2),*: 60-64.

Case, S. M., & Swanson, D. B. (1989, April). *Evaluating diagnostic pattern: A psychometric comparison of items with 15, 5, and 2 options.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Case, S. M., & Swanson, D. B. (1993). Extended-matching items: A practical alternative to free-response questions. *Teaching and Learning in Medicine, 5*, 107-115.

Case, S. M., Swanson, D. B., & Stillman, P. L. (1988). *Evaluating diagnostic pattern recognition: The psychometric characteristics of a new item format.* Paper presented at the Annual Conference on Research in Medical Education, Washington, DC.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed; pp. 201-219). New York: American Council on Education.

Dillon, G. F., Henzel, T. R., Klass, D. J., LaDuca, A., & Peskin, E. (1993, April). *Presenting test items clustered around patient cases: Psychometric concerns and practical implications for a medical licensing program.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Downing, S. M. (1992). True-false, alternate-choice, and multiple-choice items. *Educational Measurement: Issues and Practice, 11*, 27-30.

Downing, S. M., Grosso, L. J., & Norcini, J. J. (1994, April), *Multiple true-false items: Validity in medical specialty certification.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Ebel, R. L. (1972). *Essentials of educational measurement.* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement (5th ed.).* Englewood Cliffs, NJ: Prentice-Hall.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*(2), 175-186.

Fine, S. A. (1986). Job analysis. In R. E. Berk (Ed.), *Performance assessment: Methods and applications (pp. 53-81).* Baltimore: Johns Hopkins University Press.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice, 11*(4), 21-26.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement, 19*, 99-105.

Haladyna, T. M. (1991). Generic questioning strategies for linking teaching and testing. *Educational Technology: Research & Development, 39*, 73-81.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice, 11*, 21-25.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items.* Hillsdale, NJ: Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*, 999-1010.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*, 37-50.

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

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: State of the art (pp. 80-123).* Baltimore: Johns Hopkins Press.

Kaplan, R. M. (1992). Scoring natural language free-response items—A practical approach. Proceedings of the 33rd Annual Conference of the Military Testing Association (pp.514-518).

Kreiter, C. D., & Frisbie, D. A. (1989). Effectiveness of multiple true-false items. *Applied Measurement in Education, 2*, 207-216.

LaDuca, A. (1980). The structure of competence in health professions. *Evaluation & the Health Professions, 3*, 253-288.

LaDuca, A. (1994). Validation and professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation & the Health Professions, 17*, 178-197.

LaDuca, A., & Engel, J. D. (1994). On the neglect of professions theory in professions education. *Professions Education Researcher Quarterly, 15* (4), 8-11.

LaDuca, A., Taylor, D. D., & Hill, I. K. (1984). The design of a new physician licensure examination. *Evaluation & the Health Professions, 7*, 115-140.

LaDuca, A., Templeton, B., Holzman, G. B., & Staples, W. I. (1986). Item-modelling procedure for constructing content-equivalent multiple choice questions. *Medical Education, 20*, 53-56.

Lord, F. M. (1944). Reliability of multiple-choice tests as a function of number of choices per item. *Journal of Educational Psychology, 35*, 175-180.

Lord, F. M. (1977). Optimal number of choices per item—A comparison of four approaches. *Journal of Educational Measurement, 14*, 33-38.

Maatsch, J. L., Huang, R. R., Downing, S. M., & Munger, B. S. (1984). The predictive validity of test formats and a psychometric theory of clinical competence. *Proceedings of the 23rd Conference on Research in Medical Education.* Washington, DC: Association of American Medical Colleges.

Madaus, G. F. (1992). An independent auditing mechanism for testing. *Educational Measurement: Issues and Practice, 11*, 26-31.

Martinez, M. E. (1993). Problem-solving correlates of new assessment forms in architecture. *Applied Measurement in Education, 6* (3), 167-180.

Martinez, M. E., & Bennett, R. E. (1992). A review of automatically scorable constructed-response item types for large-scale assessment. *Applied Measurement in Education, 5*, 151-169.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing.* New York: Academic Press.

Schmeiser, C. B. (1992). Ethical codes in the professions. *Educational Measurement: Issues and Practice, 11,* 5-11.

Shea, J. A., Poniatowski, P. A., Day, S. C., Langdon, L. O., LaDuca, A., & Norcini, J. J. (1992). An adaptation of item modelling for developing test item banks. *Teaching and Learning in Medicine, 4* (1), 19-24.

Smith, M. L. (1991). Meanings of test preparation. *American Educational Research Journal, 28,* 521-542.

Veloski, J. J., Rabinowitz, H. K., & Robeson, M. R. (1993). A solution to the cuing effects of multiple choice questions: The un-Q format. *Medical Education, 27,* 371-375.

Chapter 5 Appendix 1

EXAMPLES OF OBJECTIVES-BASED ITEMS

Encounter: Diabetes mellitus
Objective: Recognizes new signs and symptoms in patient with established diagnosis

A 55-year old man has had insulin-dependent diabetes mellitus for most of his life. He is in the hospital recovering from a gastrointestinal operation and he is receiving regular insulin on a sliding scale. He has no glycosuria, but he has persistent ketonuria. The most appropriate management is to

(A) increase the dose of insulin
(B) decrease the dose of insulin
(C) increase his caloric intake
(D) decrease his caloric intake
(E) substitute an oral hypoglycemic drug

Encounter: Diverticula of intestine
Objective: Knows to counsel patient or family regarding current and future problems or self-care

A 34-year-old woman who is otherwise asymptomatic had an upper gastrointestinal roentgenographic study because of a 6-month history of abdominal pain. A duodenal diverticulum was found. She should be advised that

(A) the duodenal diverticulum is the cause of her pain
(B) the duodenal diverticulum should be removed surgically
(C) the duodenal diverticulum will cause gallstones
(D) long-term treatment with tetracycline will be initiated
(E) no treatment is necessary for the duodenal diverticulum

Encounter: Various diseases of the gallbladder
Objective: Recognizes new signs and symptoms in patient with established diagnosis and adjusts therapy

A 50-year-old woman, who is scheduled for elective cholecystectomy, has been taking eight aspirin tablets daily for pain caused by arthritis. In preparing for the operation, it would be best to

(A) give her a 4-donor platelet pack on the morning of the operation
(B) operate, but have platelets available if bleeding occurs during the operation
(C) discontinue her aspirin therapy and wait 2 weeks before proceeding with the operation
(D) discontinue her aspirin therapy and wait 24 hours before proceeding with the operation
(E) give the patient fresh-frozen plasma if bleeding occurs during the operation

Encounter: Osteoarthritis and allied conditions
Objective: Interprets laboratory or diagnostic studies as to underlying pathophysiology

A 73-year-old woman who has degenerative joint disease develops pain and swelling in her left knee. An x-ray film of the knee shows a narrowed joint space and linear calcifications within the joint space. The most likely finding in the joint fluid will be

(A) decreased serum glucose concentration
(B) gram-negative organisms
(C) leukocyte count > 100,000 mm^3
(D) negatively birefringent (needle-shaped) crystals
(E) positively birefringent (rhomboid) crystals

Encounter: Gout
Objective: Interprets results of diagnostic studies as to the impact on diagnosis or management

A 41-year-old man has an acute attack of gout involving his right great toe. He had one attack 8 months ago, but he has not been taking any medication. An x-ray film of the affected area would most likely show

(A) calcification of cartilage
(B) sharply marginated bone erosions
(C) subchondral osteopenia
(D) subperiosteal bone resorption
(E) no abnormality

Encounter: Prostate gland
Objective: Knows to counsel patient or family regarding current and future problems or risk factors

Two days ago, a 69-year-old man had a suprapubic prostatectomy during which 85 g of hyperplastic tissue were easily enucleated. Microscopic examination shows a 2-mm focus of adenocarcinoma. In addition to providing supportive care, he should be advised that he will also benefit from

(A) no further specific therapy
(B) total prostato-seminal-vesiculectomy
(C) hypophysectomy
(D) orchiectomy
(E) estrogen therapy

Chapter 5 Appendix 2

SELECTED EVALUATIVE OBJECTIVES AND ASSOCIATED LEAD-INS

History-Taking

Recognizes physician's best choice of words or interprets patient's own words

The best opening question is
The most appropriate initial question would be
The (physician's) most appropriate response would be

Interprets elicited history; vignette description is limited to history information

The most likely explanation (of presented case history) is

Physical Exam

Knows appropriate directed physical exam or required technique

During the physical examination, particular attention/special consideration should be given to
The physical examination should specifically focus on
The physical examination should be directed toward

Using Diagnostic Aids

Selects appropriate routine laboratory or diagnostic studies (study of choice, usually initial)

The most appropriate initial diagnostic study is
At this time, the most appropriate diagnostic study/procedure is
The best initial diagnostic step/study is
The most appropriate next step is to (response options list diagnostic studies)

Evaluates utility of diagnostic and invasive, special, non-routine studies

NOTE: The studies of choice are usually follow-up and more invasive than initial studies (e.g., biopsies). Results of prior diagnostic studies are usually described in the stem.

The most reliable next diagnostic test is
The most appropriate next step is (response options list, invasive diagnostic studies)

Making Diagnosis & Defining Problems

Selects most likely diagnosis or evaluates differential in light of history and/or physical and/or diagnostic test findings

The most likely diagnosis is (given diagnostic vignette in stem)
These findings are most likely a result of (response options are diagnoses)

Interprets vignette and identifies the indicator for consultation or further diagnostic assessment (Response options are indications)

Which of the following findings should prompt referral to a (specialist)?
In this patient, which of the following requires consultation with a specialist?
Further diagnostic assessment is mandated by
The most important indication for consultation (with a particular specialist) is the presence of

Assesses severity of patient condition and makes judgment as to current status, prognosis or need for further action (Response options are inferences or conclusions referenced to the patient in complete sentences)

At this time it is most appropriate to conclude that
The most accurate statement concerning the patient is
The most likely explanation for this patient's worsening condition is

Managing Therapy

Knows priorities for, or immediate consequences of, selecting among various interventions or therapies

Priorities in management include
(Therapy/intervention) will be appropriate for this patient if/when
The most appropriate next step is (response options focus, for example, on whether to obtain more details of the history or physical or order more studies or observe or begin treatment)

Knows indications (based on signs and symptoms) for immediate medical intervention (Emergency situations)

The most appropriate immediate management would be to

Knows appropriate present management of selected conditions (excludes all-drug options); often "wait and see" or other benign intervention

At this time, the most appropriate management is to
The most appropriate initial management is to
The most appropriate next step is to (response options are management-oriented, not diagnostic studies)

Recognizes indications for use of medications or prophylactic drugs or vaccines (e.g., drug of choice)

The most appropriate pharmacotherapy (for specific patient) is
In managing a patient with (condition), the medication most appropriate is

Knows indications for hospital admission or other appropriate setting, including moving patient to ICU, CCU

The factor most influential in deciding if the patient should be admitted to the hospital/special care unit is
The most appropriate next step is to (correct response option is to admit the patient to the hospital or special care unit)

Knows importance of educating patient or family regarding self-care, therapeutic regimen (e.g., BP measurement, home glucose monitoring) (Focus is on behavior regarding the specified therapy)

The patient (receiving a specific medication/therapy) should be told to avoid/be told to expect/be warned about
The patient should be told to/advised to (response options include, for example, home blood glucose measurement, self-examination)

Chapter 5 Appendix 3

SAMPLE CASE CLUSTER

A 45-year-old nurse sticks herself with a needle after it was used to draw blood from a 35-year-old jaundiced patient. The nurse is in good health when she comes to your office for a work-up of the incident. She takes only lovastatin for hyperlipidemia. Her last tetanus toxoid injection was 8 years ago. Laboratory studies done on the nurse and patient show:

TESTS	NURSE	PATIENT
Serum		
AST, GOT	16 U/L	450 U/L
ALT, GPT	8 U/L	560 U/L
Alkaline phosphatase	50 U/L	200 U/L
Serologies		
HbsAg	Negative	Positive
Anti-HBc	Negative	Negative
Anti-HAV (IgM)	Negative	Negative
Anti-HAV (IgG)	Negative	Positive
HIV	Negative	Negative

1. Other persons who should be tested are:
   - (A) the nurse's household contacts
   - (B) the other emergency department staff who were exposed to the patient
   - (C) the patient's child's playgroup
   - (D) the patient's household contacts
   - (E) no one else needs to be tested
2. The nurse should receive
   - (A) hepatitis B vaccine
   - (B) hyperimmune B globulin
   - (C) hyperimmune B globulin and hepatitis B vaccine
   - (D) immune serum globulin
   - (E) tetanus toxoid
3. The patient should receive
   - (A) hepatitis B vaccine
   - (B) hyperimmune B globulin
   - (C) hyperimmune B globulin and hepatitis B vaccine
   - (D) immune serum globulin
   - (E) none of these

The nurse and patient are treated appropriately. Two weeks later the nurse develops right upper quadrant pain, low-grade fever, and dark urine.

4. The LEAST likely explanation for her symptoms is
   (A) hepatitis A
   (B) hepatitis B from the needle-stick contact with the patient
   (C) hepatitis C
   (D) gallbladder disease
   (E) reaction to lovastatin

The nurse admits to heavy intake of alcohol. Testing shows no other abnormalities and her symptoms resolve with abstinence from alcohol. Six months later she has a routine examination as part of an application for life insurance coverage. She is asymptomatic. Laboratory test results are:

Serum	
AST, GOT	100 U/L
ALT, GPT	110 U/L
Alkaline phosphatase	100 U/L
Bilirubin, total	1.0 mg/Dl

5. Which of the following statements concerning these findings is correct?
   (A) Her lack of symptoms is a favorable prognostic sign
   (B) It is unlikely that she has chronic hepatitis because she is female
   (C) These values are expected as a consequence of her history of alcohol ingestion
   (D) The results represent a laboratory error
   (E) The results are most likely an early sign of AIDS

Repeat testing done the next day shows the following:

HBsAg	Negative
Anti-HBc	Positive
Anti-HAV (IgG and IgM)	Negative

6. Based on these findings, the most appropriate next step is
   (A) administration of immune serum globulin to her family members
   (B) administration of hyperimmune B globulin
   (C) liver biopsy
   (D) repeat liver chemistry profile in 6 months
   (E) test for antibodies to smooth muscle

Chapter 5 Appendix 4

SAMPLE AMPLIFIED OBJECTIVE

Evaluative Objective

Assesses severity of patient condition and makes judgment as to current status, prognosis, or need for further action.

Sample Item

Encounter: Cranial or ocular injury
Objective: Assesses severity of patient condition

A 55-year-old woman, who is an established patient, has been returned to the office by her adult son because of continuing complaints following an auto accident. At that time she suffered severe laceration when she was hit in the occipital skull by a piece of metal. For the past six weeks she has complained of headaches and she has had difficulty seeing. During this period, her family has noticed that she is behaving strangely. She does not seem to recognize objects even though her vision appears to be intact. She is forgetful, especially of recent events. She appears somewhat indifferent to friends and family members, and is described as "socially inappropriate." The greatest concern is that this patient

(A) has experienced an exacerbation of the occipital injury
(B) has experienced a major psychiatric illness, with the experience of the auto accident as a precipitating factor
(C) has had a major bilateral stroke in the anterior cerebral arteries
(D) has suffered damage to the anterior temporal lobes and frontal lobes in her initial auto accident
(E) is having episodes of atrial fibrillation or other cardiac problems

Answer: D

General Description

Given a description of an existing clinical problem or condition in a specific patient, the examinee will assess severity of illness by making appropriate judgments about clinical status, prognosis, or therapeutic options.

Faceted General Description

A
Given a {description of an existing clinical problem or condition}

B
in a {*specific patient*}, *the examinee will make a judgment about appropriate*
C *D* *E*
{*clinical status*}, or {*prognosis*}, or {*therapeutic options*}.

Content Limits

A:{description of an existing clinical problem or condition}
Use clinical problems/conditions in appropriate domain reference.

B:{specific patient}

b_1: adult, black, female
b_2: adult, white, male
b_3: elderly, black, male
b_n: age, race, sex

C:{clinical status}

c_1: admission to the hospital is required
c_2: specific infectious agent is responsible
c_3: no further follow-up is required

D:{prognosis}

d_1: patient is at risk for _____
d_2: the most likely consequence will be _____
d_3: the complication most likely to arise is _____

E:{therapeutic option}

e_1: surgical valve replacement will be required
e_2: serology is essential for further evaluation
e_3: no change in pharmacotherapy is needed
e_4: referral to _____ is needed

Response Limits

1. Use 4-, or 5-option, A-type MCQ preferably.
2. Response options are declarative sentences stating various assessments of severity.
3. Response options may be drawn from ONE facet (e.g., C or D or E), or from SEVERAL facets (e.g., one each from C and D and E).
4. Correct therapeutic option responses need only be preferable to incorrect responses.