Chapter 7

Written Tests: Constructed-Response and Selected-Response Formats

Steven M. Downing

Contact Information:

Steven M. Downing, PhD
Associate Professor
University of Illinois at Chicago
College of Medicine. Department of Medical Education (MC 591)
808 South Wood Street, Office 986-C
Chicago, Illinois 60612-7309
Phone: 312.996.6428
Fax: 312.413.2048
E-mail: sdowning@uic.edu

Acknowledgments

The author is most grateful to Thomas M. Haladyna, PhD for his review of and constructive criticisms and suggestions for this chapter.

****

Downing, S.M. & Yudkowsky, R. (Eds.) ***Assessment in Health Professions Education*** (In Press). LEA. Please do not cite or quote this chapter without the explicit written permission of the authors.

Introduction

The purpose of this chapter is to provide an overview of the two written testing formats most commonly utilized in health professions education: the constructed-response (CR) and the selected-response (SR) item formats. This chapter highlights some key concepts related to the development and application of these testing modalities and some of the important research evidence concerning their use. This chapter is not intended to be a complete item writing guide or a comprehensive and in-depth critical review of the current theoretical and research literature on written testing or a scholarly defense of written testing in either modality.   Rather, the objective of this chapter is to provide a practical summary of information about developing and effectively using CR and SR methods to test cognitive achievement in health professions education, with some suggestions for appropriate use.

CR and SR Formats

The generic terms constructed-response (CR) and selected-response (SR) are accurately descriptive of how these two testing formats work. CR items require the examinee to produce a written response to a stimulus, usually a question or a statement. In this chapter, CR items are discussed as direct or implied open-ended questions or other types of stimuli that require examinees to write (or type) responses or answers, which are then read and scored by content-expert human judges or raters. Essay tests are the most common application of the CR item form in health professions education. Such a narrow definition of CR tests – limited to essay questions alone – would be disputed by many educational measurement professionals who view CR testing as a type of performance testing (e.g., Haladyna, 2004). SR items require examinees to choose a correct or best answer from a fixed listing of possible answers to a question or other stimuli. Examinee answers to SR items may be computer-scored, using answer keys (listing of correct or best answers) developed by content experts. Multiple-choice items (MCQs) are a common example of the SR item form. Table 7.1 summarizes some characteristics of each format discussed in this chapter.

INSERT TABLE 7.1 ABOUT HERE

The prototypic CR item type is the essay question. For this chapter two general types of essays are discussed – those requiring long answers and those requiring short answers. A long-answer essay may require the examinee to write 1-2 or more pages in response to the question, while short-answer essay questions may require a 1-2 paragraph written response.

The multiple-choice item (MCQ) is the prototypic SR item type. All other examples of fixed-answer test item formats may be considered a variant of the multiple-choice item type. MCQ variants include: the true-false, alternate-choice, multiple-true-false, complex MCQ, matching, and extended matching item types. Table 7.2 lists some examples.

INSERT TABLE 7.2 ABOUT HERE

Assessment Using Written Tests

What are written tests good for? Written tests are useful in the measurement of cognitive knowledge or to test learning, achievement, and abilities. Referring to Miller's Pyramid, the "knows" and "knows how" level at the base of the pyramid are best measured by written tests. And, the ACGME toolbox suggests the use of written tests for measuring cognitive knowledge (Downing & Yudkowsky, Chapter 1, this volume). Most cognitive knowledge is mediated verbally, such that humans acquire cognitive knowledge through written or spoken words or by visual, auditory or other stimuli that may be translated or mediated verbally. Thus, written tests are ideally suited to test verbal knowledge. (The nature of "cognitive knowledge" and its acquisition is far beyond the scope of this book.) Many educational measurement texts discuss high-inference and low-inference written item formats, to distinguish the assessment of more abstract verbal knowledge from more concrete verbal knowledge (e.g., Haladyna, 2004; Linn & Miller, 2005).

Written assessments are best suited for the assessment of all the types of learning or cognitive knowledge acquired during courses of study in the health professions – through curricula delivered in classrooms, textbooks, lectures, library and internet research, student discussions in small learning groups, problem-solving group activities, on-line teaching/learning environments, and so on. Written tests are most often and most appropriately used to assess knowledge acquisition – as formative or summative assessments, to provide feedback on learning or to measure the sufficiency of learning in order to proceed in the curriculum. Written tests are not at all useful to test performance or "doing," unless that performance happens to be the production of writing (which can be tested only by written tests).

The primary guiding factor in determining the appropriateness of any testing format relates to its purpose, the desired interpretations of scores, the construct hypothesized to be measured, and the ultimate consequences of the test. The characteristics of the testing format should match the needs for validity evidence for some particular assessment setting and there should be a clear rationale for choice of the written format, given the validity needs of the assessment. For example, if the goal is to test student cognitive knowledge about the principles of effective patient communication or the understanding of various principles of effective communication with patients, a written test may match the purpose of the test and the required needs for specific types of validity evidence to support score inferences.   But, in order to measure students' use of communication skills with patients requires some type of performance test – a simulation, a standardized oral exam, or a structured observation of student communication with patients in a real setting. A written test would be mismatched to the purpose of this test and the required validity evidence, given the intended purpose of the test.

Both the CR and the SR have some unique strengths and limitations, as noted in Table 7.1.   Both testing formats have been researched and written about for nearly a century. Strong beliefs, long-held traditions, and vigorous opinions abound.   In this chapter, we review some of the science and research evidence and summarize the best practice that follows from this research.

Constructed-Response Items

Constructed-response (CR) items, in some form, have been used to test students for centuries. In this chapter, CR items are discussed only as essay questions – either short- or long-answer essay questions.

CR formats have many strengths. For instance, the CR format is the only testing format useful for testing writing skills such as the adequacy of sentence and paragraph construction, skill at writing a persuasive argument, ability to organize logical thoughts, and so on. All CR items require non-cued written answers from examinees. The CR item format may permit the essay reader to score specific steps in working through a problem or the logic of each step used in reasoning or problem solving, which may facilitate partial credit scoring (as opposed to “all or nothing” scoring). CR formats may be most time efficient (for the instructor) in testing small groups of students, since less time will be spent writing essay questions or prompts than in creating effective SR items. Small groups of examinees also may make the essay scoring task time efficient. And, essay questions are usually easier to write than MCQs or other SR formats.

However, there are also many issues, challenges, and potential problems associated with essay tests. CR tests are difficult to score accurately and reliably. Scoring is time consuming and costly. Content-related validity evidence is often compromised or limited, especially for large content domains, because of sampling issues related to testing time constraints. And, there are many potential threats to validity for CR items, all related to the more subjective nature of essay scores and various biases associated with human essay readers. There are fewer psychometric quality-control measures, such as item analysis, available for CR items than for SR items.

The purpose of the CR test, the desired interpretation of scores, and hypotheses about the construct measured – validity – should drive the choice of which written format to use in testing cognitive knowledge. For instance, if the goals and objectives of instruction relate to student achievement in writing coherent explanations for some biochemical mechanism and in tracing each particular stage of its development, an essay test may be a good match. “Writing” is the key word, since only CR item forms can adequately test the production of original writing. (SR formats can test many of the components of writing, such as knowledge of vocabulary, sentence structure, syntax and so on.)

Anatomy of a Constructed-Response Prompt

CR items or questions are often referred to generically as prompts, since these stimuli can take many forms in performance testing: written questions, photographs, data tables, graphs, interactive computer stimuli of various types, and so on. These general stimuli serve to prompt a CR response, which can then be scored. In this chapter, we discuss CR items as essay questions only, since these are the most frequently used type of CR format in health professions education worldwide.

An essay question or prompt consists of a direct question on a specific focused topic and provides sufficient information to examinees to answer the question. All relevant instructions concerning answering the question, such as expected length of answer, time limits, specificity of answer, and so on must be clearly stated. See Table 7.2 for some examples.

Basic Principles of Writing Constructed-Response Items

"Writers of performance assessment items must adhere to the same rules of item writing used in the development of multiple-choice test items." (Welch, 2006, p. 312.) Table 7.3 presents these item writing principles, as defined by the educational measurement textbooks and the empirical research on these principles (Haladyna, Downing, & Rodriguez, 2002).

CR item writing benefits from attention to these principles and revisions and editing based on independent review by other content experts (Downing, 2006). Clarity of meaning is an essential characteristic for all test items, since such text is highly scrutinized by examinees for subtle meaning.   As in all testing, the content to be tested is the most fundamental consideration; the format selected for the test is always of secondary importance.

During the preparation of the essay-type question, a model or ideal answer to the question should also be prepared by the author of the question, just as a correct or best answer key should be designated by a SR item author. The specificity of the model answer must match the directions to examinees. This model or ideal answer will form the basis of a scoring rubric (the scoring key for CR items) used in the actual scoring of the response to the essay question (see Table 7.4 for example).

The CR item, including its specific directions for examinees, the ideal answer, and the actual scoring rubric should be prepared well in advance of the test administration, so that time for review, revision and editing is available.

Short-Answer versus Long-Answer Constructed-Response

Short-answer CR items require answers of a few words, a few sentences, or a few paragraphs, whereas long-answer CR items require written responses of several pages in length. The purpose of the assessment and the content-related validity requirements for broad sampling versus depth of sampling should drive decisions about CR length.   In achievement assessment for most classroom settings, breath of sampling is important because the purpose of the test is to generalize to an examinee’s knowledge of some large domain of knowledge from a limited sample.   If CR tests are used, short-answer essays permit broader sampling of content than long-answer essays, because more questions can be asked and answered per hour of testing time.

If the purpose of the test is to sample a narrow domain of knowledge in great depth, long-answer essays may be the most appropriate format. Long-answer essays permit asking an examinee to produce answers of great detail, probing the limits and depths of knowledge about a single topic or content area. In some cases, long-answer essays may be appropriate, but generally these longer essays are poor samples of large domains and therefore lack generalizability and validity evidence.

Scoring Constructed-Response Items

Scoring is a major validity challenge for CR items. CR scoring is inherently subjective and therefore requires attention to a number of issues in order to reduce the negative effect of subjectivity on scoring validity. In this context, we discuss scoring methods and rater characteristics together with some basic recommendations to increase scoring accuracy.

CR Scoring Methods

There are two different approaches to essay scoring: analytic or holistic ratings. In analytic methods, essays are rated in several different categories or for several different characteristics. For example, analytic scoring might require ratings of the accuracy of the answer to the question and the specificity of the answer, the organization of the written answer, the writing quality, and so on. Analytic methods require the rater to concentrate on several different aspects of the essay, all of which are presumably related to the quality of the essay answer and the construct intended to be measured by the essay question. Score points are assigned to each analytic segment or aspect of the essay, based on some rationale. Holistic or global ratings require the essay reader to make only one single rating of the overall quality of the written answer.

Which is the best method, analytic or holistic? The answer depends on the purpose of the CR test. Analytic scoring methods may permit feedback to examinees on more specific aspects of performance than do global methods. However, many of the separate characteristics rated in the analytic method may correlate highly with each other, thus reducing the presumed benefit of analytic scoring methods. Global or holistic ratings are generally more reliable than individual ratings, but the intended use of the CR rating data should be the major factor in deciding on analytic or holistic methods (See McGaghie et al, Chapter 8, this volume). Analytic methods usually require more scoring time than global methods, so feasibility and practicality will also be a factor in the choice of method.

Analytic methods may permit the weighting or differential allocation of partial credit scores somewhat easier or more logically than global methods. For an essay item in which several different essay traits are rated, it is possible to allocate the total score for the essay differentially across the rating categories. For example, the content and structure of the essay answer may be weighted more highly than the writing quality and the organization of the answer; score points for the answer would be allocated accordingly. Analytic methods may assist the essay reader in staying focused on the essential features of the answer.

Model Answers

Whichever scoring method is used, an ideal or model answer should be prepared for each essay question rated. This model answer should list all of the required components to the answer. Model answers are analogous to the scoring key for a SR test, so they should be reviewed by content experts for accuracy and completeness. Model answers strive to reduce the subjectivity due to human raters, by introducing some objectivity and standardization to the scoring process.

Essay Raters

Human readers or raters of essay answers are essential. The subjectivity of human raters creates a potential major scoring problem. Human raters bring biases and many other potential sources of rater error to the task, so counterbalancing efforts must be taken to try to reduce problems due to rater subjectivity.

It is recommended that two independent raters read every essay answer and that their separate ratings be averaged—especially for essays that have higher stakes or consequences for examinees. The expectation is that averaging the ratings from two independent readers will reduce bias. For example, if one rater tends to be a “hawk” or severe and the other rater tends to be a “dove” or lenient, their mean rating will offset both the severity and the leniency bias. On the other hand, if both raters are severe or both are lenient, the average rating will do nothing to offset these rating errors or biases and will, in fact, compound the problem.

It is also often suggested that essay raters read all the answers to one essay question for all examinees, rather than reading all answers to all questions for a single examinee. It is thought that essay raters do better if they can focus on one essay answer at a time, but there is little evidence to support this recommendation.

Scoring Rubrics

A scoring rubric is a detailed guide for the essay rater and attempts to reduce some of the inherent subjectivity of human raters by stating pre-specified behavioral anchors for ratings. Scoring rubrics can take many forms in providing anchors and specific detail for the scoring task; the specific forms will differ for analytic or global rating methods. See Table 7.4 for a simple example of an analytic scoring rubric, to be used in the scoring of a short essay answer. Note that the use of essay scoring rubrics fits well with the recommendation to use model answers and two independent raters—all suggestions intended to reduce the idiosyncratic subjectivity due to human raters.

INSERT TABLE 7.4 ABOUT HERE

Threats to Validity of CR Scoring

Both the content underrepresentation (CU) and the construct-irrelevant variance (CIV) validity threats are potential issues for CR tests (Downing, Chapter 2, this volume; Downing & Haladyna, 2004; Haladyna and Downing, 2004; Messick, 1989). For example, if only long-answer essays are used for classroom-type achievement assessment, content underrepresentation is a potential threat, especially for large achievement domains. Long-answer essays may undersample large domains, since only a few questions can be posed and answered per hour of testing time.

Construct-irrelevant variance (CIV) threats to validity abound in essay-type testing. Rater error or bias due to reader subjectivity is the greatest source of potential CIV for essay testing. Unless great care is taken to reduce or control this type of rater error, collectively known as rater severity error, components of the final score assigned to essay answers can be composed of reliable ratings of irrelevant characteristics. Raters are notoriously poor, even when well trained, at controlling their tendencies to assign biased scores to essays.

The well known rater errors of halo, leniency, severity, central tendency, and idiosyncratic rating fully apply to essay readers (McGaghie et al, Chapter 8, this volume). Tracking of raters and providing frequent feedback to essay raters on their performance, especially relative to their peers, may help temper some of these CIV errors. And using the average rating of two independent raters, who have different biases, may diminish some of the ill effects of rater bias. Obviously, formal written model answers seek to lessen the subjectivity of ratings, as do the use of written scoring rubrics.

Another source of error concerns examinee bluffing, which is sometimes attempted by examinees who do not know the specific answer to the question posed. Some bluffing methods include: restating the question to use up required space; restating the question in such as way as to answer a different question; writing correct answers to different questions (which were not posed in the prompt); writing answers to appeal to the biases of the essay reader, and so on (e.g., Linn & Miller, 2005). If bluffing attempts are successful for the examinee, CIV is added because the scores are biased by assessment of traits not intended to be measured by the essay.

Other potential CIV issues relate to the quality of handwriting, which can be either a positive or negative bias, writing skill (when writing is not the main construct of interest); skill in the use of grammar, spelling, punctuation (when these issues are not the primary construct); and so on. All such extraneous characteristics of the written response can unduly influence the essay reader, in either a positive or a negative manner, adding CIV to the scores and thereby reducing evidence for validity.

Constructed-response Format: Recommendations and Summary

The constructed-response (CR) format is good for testing un-cued written responses to specific questions. If the purpose of the assessment is to test student achievement of the relevant content in a written form – where components of writing are critical to the content – CR is the format of choice. The CR format is the only format to use to test the actual production of writing.

CR formats may be preferred if the number of examinees is small, since scoring essay responses may take less time than writing selected-response items. Also, it may be possible to assign partial credit to CR answers in a logical or defensible manner. Short-answer essays are preferred to long-answer essays for most classroom achievement assessment settings, because of the possibility for better content-related validity evidence.

Scoring of essay answers are a challenge, due to the inherent subjectivity of the human essay reader. Using at least two independent and well trained raters, who use model or ideal answers and clear scoring rubrics to anchor their scores, is recommended. The provision of specific and timely feedback to essay raters may help to reduce some rater bias. The choice of analytic or global and holistic methods of scoring depends on the purpose of the test, the content of the essays, the stakes associated with the scores and feasibility issues.

Selected-Response Items

> "*Any* aspect of cognitive educational achievement can be tested by means of either the multiple-choice or the true-false form."
>
> (Ebel, 1972, p. 103)

This quote from Robert L. Ebel, a scholar of the SR format, provides an appropriate introduction to this section.

Selected-response (SR) items, typified by multiple-choice items (MCQ) as the prototypic form, are the most useful written testing format for testing cognitive knowledge in most health professions education settings.   Some examples of commonly used SR item forms are presented in Table 7.2.

The SR item format was developed nearly a century ago to provide an efficient means of cognitive testing for large-groups of examinees. Ebel (1972) presents a brief history of the early development of the MCQ format and its first major use by the U.S. military for recruit selection testing in the early twentieth century. In discussions of the relative merits of SR and CR testing, it may be instructive to remember that SR formats were introduced to overcome shortcomings of the CR format.

MCQs are useful for testing cognitive knowledge, especially at higher levels. MCQs are most efficient for use with large groups of examinees because the time spent in preparing test items prior to administering the test is generally less than the time required to read and score CR items after the test, because MCQs can be easily and rapidly computer scored. Effective MCQs can be re-used on future tests, if stored securely in a retrievable item bank. Also, MCQs are most efficient for testing large knowledge domains broadly, so that the test is a representative sample of the total content domain, thus increasing the content-related validity evidence, and permitting valid inferences or generalizations to the whole of the content domain. MCQs can be scored accurately, reliably, and rapidly. Meaningful MCQ score reports – providing feedback to students on specific strengths and weaknesses – can be produced easily by computer and in a timely and cost effective way – thus potentially improving the learning environment for students. Sound psychometric theory, with a large research base and a lengthy history, underpins MCQ testing. Validity and reliability theory, item analysis and other test quality-control methods, plus an emerging theory of MCQ item writing – provide support for the use of well crafted MCQs in the testing of cognitive achievement (Downing, 2002a, 2006; Downing & Haladyna, 1997).

For a complete and in-depth scholarly treatment of the MCQ format and its variants, refer to Developing and Validating Multiple-Choice Test Items, third edition (Haladyna, 2004). This book-length treatment is the best single source of current research on the MCQ form and its application in educational testing.

Anatomy of an MCQ

A multiple-choice item consists of a *stem* or lead-in, which presents a stimulus or all the necessary information required to answer a direct or implied *question*. The stem and question are followed by a listing of possible answers or *options*.

Basic Principles of Writing Effective MCQs

Over many years of development, research and widespread use, principles for creating effective and defensible MCQs have emerged. These evidence-based principles have been summarized by studies, which reviewed the advice to MCQ item writers by authors of the major educational measurement textbooks and the recommendations based on relevant empirical research concerning these item writing principles (Haladyna & Downing, 1989 a, b; Haladyna, Downing, & Rodriguez, 2002).   Table 7.3 lists a summary of these 31 principles and is adapted from Haladyna, Downing, and Rodriguez (2002).

There are empirical studies supporting about one-half of these 31 principles of effective item writing and most major educational measurement textbook authors endorse most of these principles. Thus, these 31 principles offer the best evidence in practice for creating effective and defensible MCQs. But, these general principles alone are not sufficient to assist the MCQ item writer in creating effective test items. For a excellent and detailed item writing guide, aimed specifically toward the health professions educator, see Case and Swanson (1998) and the National Board of Medical Examiners (NBME) website (www.nbme.org). This item writing guide presents excellent suggestions and many relevant examples of effective and ineffective SR items.

MCQs which violate one or more of these standard item writing principles have been shown to disadvantage some students. In one study, flawed items were artificially more difficult for medical students and misclassified 14 percent of students as failing the test when they passed the same content when tested by non-flawed MCQs (Downing, 2005).

Overview of Principles for Effective MCQs

The most effective MCQs are well focused on a single essential or important question or issue. The single most important requirement is that the item's content is relevant, important, and appropriate. Most of the information needed to answer the question is contained in the stem of the item, which is worded positively, and concludes with a direct (or indirect) question. Options (the set of possible answers) are generally short, since most of the information is contained in the stem of the item. There is a good match between the cognitive level posed by the question and the instructional objective of the instruction. Generally, many items test higher-order cognitive objectives of instruction (such as understanding, application, evaluation) using novel content; few items test the lower levels of the cognitive domain such as recall and recognition. The set of options are homogeneous such that all possible answers are of the same general class and every option is a plausible correct answer. One and only one of the options is the correct (or best) answer to the question posed in the stem. Experts agree on the correct or best answer. The wording of the MCQ is extremely clear so that there are no ambiguities of language. No attempt is made to deliberately trick knowledgeable examinees into giving an incorrect answer. All clues to the correct answer are eliminated from the item, as are all unnecessary complexities and extraneous difficulty, and all other ambiguities of meaning (Baranowski, 2006). The MCQ is drafted by the item author – an expert in the content of the item – who asks another content expert to review the draft item and its form. Sufficient time is allowed for review comments to be considered and changes to be incorporated into the final item (Downing, 2006).

On the other hand, a poorly crafted or flawed MCQ may test trivial content, at a low level of the cognitive domain (recall or recognition only). The item may have an unfocused stem, so that the question is not clearly stated – so that the examinee must read all of the options in order to begin to understand the question. Such a flawed MCQ may be worded negatively and so ambiguously that examinees are confused about the question being asked. The stem may be a non-focused, open-ended statement that requires the examinee to read all the options first in order to understand what question is being asked. There may be no correct or best answer to the question or more than one correct answer, so that the correctness of the scoring key can not be defended.   The flawed MCQ may incorporate inadvertent cues to the correct answer, so that uninformed examinees can get the item correct; or, the item may be so ambiguously written that examinees who actually know the content intended to be tested by the MCQ get the item wrong (Downing, 2002a).

Elimination of five common flaws in MCQs may greatly reduce the ill effects of poorly crafted MCQs. These flaws are: unfocused stems, negative stems, the “all of the above” and the “none of the above” options, and the so-called partial-K type item (Downing, 2005). This study and others (e.g., Downing, 2002b) suggest that classroom achievement tests in the health professions typically utilize many flawed items – up to about one-half of the items studied had one or more item flaws, defined as a violation of one or more of the 31 evidence-based principles of effective item writing. And, these item flaws negatively impacted student achievement measurement and biased pass-fail decisions made from scores of tests composed of flawed items.

Creative Item Writing

Writing effective MCQs is both art and science. The *science* is provided by the evidence-based principles noted in Table 7.3. The *art* is associated with variables such as effective item writer training, use of effective training materials, practice, feedback, motivation, item review and editing skills, writing ability and so on. Writers of effective MCQ items are trained not born. Content expertise is the single most essential characteristic of an effective item writer. But content expertise alone is not sufficient, since item writing is a specialized skill and, like all skills, must be mastered through guided practice and feedback on performance. There is no reason to suspect, for example, that an internationally recognized expert in some health sciences discipline will necessarily be an expert MCQ item writer, unless that individual also has some specialized training in the science and art of item writing.

The world is awash in poorly written MCQs (Downing, 2002c). Writing effective, creative, challenging MCQs – which test important knowledge at higher levels – is a difficult and time consuming task. Lack of time for already overburdened instructors may be one major reason that there are so many poorly crafted MCQs used in typical classroom tests in the health professions. But the weakness is not with the MCQ format itself; the issues result from the poor execution of the format and the consequent negative impact of such poor execution on students.

Some MCQs Issues

Many criticisms, issues, and questions arise about MCQs and the details of their structure and scoring. Some of these concerns are reviewed here, with recommendations for practice, based on the research literature.

Number of MCQ Options

Traditionally, MCQs have four or five options. The question of the optimal number of options to use for an MCQ item has been researched over many years. So, the recommendation to use a minimum of three options is based on solid research. (See table 7.3, principle # 18) . A meta-analysis of studies by Rodriquez (2005) on the optimal number of options shows that generally three options is best for most MCQs.

Most four- or five-option MCQs have only about three options that are actually selected by 5 percent or more of the examinees and have statistical characteristics that are desirable (Haladyna & Downing, 1993). Incorrect options that are selected by 5 percent or more of examinees and have negative discrimination indices are called functional distractors. (See Downing, Chapter 5, this volume, for a discussion of item analysis data and its use.)

Since few examinees typically choose dysfunctional options, the recommendation is to "develop as many effective choices as you can, but research suggests three is adequate" (Haladyna, Downing, & Rodriguez, 2002, p. 312). Using more than three options may not do much harm to the test, but will add inefficiencies for item writers and examinees and permit the use of fewer total MCQs per hour. So, the best advice is to develop as many plausible incorrect options as feasible, noting that plausibility will ultimately be determined empirically by reviewing the item analysis data showing the number of examinees who actually chose the incorrect options. The use of three-option MCQs require a sufficient number of total MCQs be used on the test – the usual advice being a minimum of about 35-40 items total. Also, note that items on a test may have varying number of options, such that some items may have three options while other items naturally have 4, 5 or even more options.

Three-option MCQ critics suggest that using fewer than 4-5 options increases random guessing and reduces test score reliability. Of course, for a single MCQ, the probability of randomly guessing the correct answer is .33 for a three-option item and .20 for a five-option item. But, this random guessing issue is not usually a problem, for well written MCQs, targeted in difficulty appropriately, and used in sufficient numbers to overcome any meaningful gain from an occasional lucky guess. On the issue of reliability, it is true that three-option items will be slightly less reliable than 4-5 option items, but this slight decrease in scale reliability is rarely meaningful (Rodriguez, 2005).

Random Guessing

Random guessing on SR items is usually overestimated and concerns about guessing may be overstated. If SR items are well written, targeted at appropriate difficulty levels, reviewed and edited to eliminate all cues to the correct answer, random guessing is usually not a major problem. Examinees may be able to get an occasional item correct using only a lucky random guess so it is important to use sufficient numbers of total SR items on the test. If items are too difficult, examinees may have no choice but to blindly guess, so using appropriate item difficulty levels is important.

Random or blind guessing differs from informed elimination of incorrect answers, in which examinees use partial knowledge to eliminate some options and narrow their selection to the correct answer. In real life, partial knowledge is frequently used to solve problems and answer questions. We rarely have complete knowledge for decision making, especially in the health professions. We do gain information about student ability or achievement even when students use partial knowledge to answer SR items.

Random guessing is not a good strategy to achieve a high or even a satisfactory score on an SR test. For example, consider a 30-item MCQ test in which each item has three-options. The probability of getting one item correct is .33 – a good chance of randomly guessing a correct answer on that single item. But, to get two items correct using chance alone, the probability falls to .11; and, to get three items correct using random guesses only, the chance falls to .04. Even for a fairly short test of 30 items, using 3-option MCQs, the probability of getting a score of 70 percent correct is only .000036. When a more typical test length of 50 items, each with 3 options, is used, the probability of getting a good score of 75 percent correct falls to .00000000070. The odds of achieving a high score on a test using random guessing alone are not good and most students understand that random guessing is not a good strategy to optimize their test scores.

The best defense against random guessing on MCQs is to create well crafted items and to present those items in sufficient numbers to reduce any major impact resulting from some random guessing.

Correction-for-Guessing Scoring

Methods or formulas used to score MCQ tests have been researched and debated for many years. There are two basic methods used to score SR tests: count the number of correct items (number-correct score) or use a formula to try to “correct” the number-correct score for presumed guessing. Test users have disagreed about using such formula scores throughout the history of SR testing.

The simple count of the number of items marked correctly is usually the best score. Raw scores such as these can be converted to any number of other metrics, such as percent-correct scores, derived scores, standard scores, and any other linear transformation of the number-correct score (Downing, Chapter 5, this volume).

All correction-for-guessing formula scores attempt to eliminate or reduce the perceived ill effects of random guessing on SR items. These formulas usually work in one of two ways: they try to reward examinees for resisting the temptation to guess or they actively penalize the test taker for guessing (Downing, 2003). However intuitively appealing these guessing corrections may be, they do not work very well and they do not accomplish their stated goals. Both the corrected and uncorrected scores correlate perfectly (unless there are many omitted answers), indicating that both scoring methods rank order examinees identically, although the absolute values of scores may differ. Further, no matter whether examinees are directed to answer all questions or only those questions they know for certain (i.e., to guess or not to guess), savvy, testwise, or bold examinees know that they will usually maximize their score by attempting to answer every question on the test, no matter what the general directions on the test state or what formulas are used to derive a score. So, corrections for guessing tend to bias scores (e.g., Muijtjens, van Mameren, Hoogenboom, Evers, & van der Vleuten, 1999) and reduce validity evidence by adding construct-irrelevant variance (CIV) to scores, because boldness is a personality trait, and not the achievement or ability construct intended to be measured by the test.

Testlets: Context-Dependent Item Sets

One special type or special use of MCQs is in the testlet or context-dependent item set (e.g., Haladyna, 1992).   See Table 7.2 for an example of a testlet. Testlets consist of stimulus materials which are used for two or more independent items, presented in sets. For example, a testlet could consist of a paragraph or two giving a detailed clinical description of a patient, in sufficient detail to answer several different questions based on the same clinical information. One item in the testlet might ask for a most likely diagnosis, another question for laboratory investigations, another on therapies, another on complications, and final question on expected or most likely outcomes.

Testlets are excellent special applications of SR or MCQ items. Testlets are efficient, in that a single stimulus (stem, lead-in) serves multiple items. Several items can be written for the common stem and, for test security purposes, different items can be used on different administrations of the test. Testlets permit a more in-depth probing of a specific content area.

Some basic principles of testlet use must be noted. All items appearing on the same test with a common stem must be reasonably independent such that getting one of the items incorrect does not necessarily mean getting another item incorrect. Obviously, one item should not cue the answer to another item in the set. Each item in the testlet is scored as an independent MCQ, but the proper unit of analysis is the testlet score and not the item score, especially for reliability analysis (Thissen & Wainer, 2001; Wainer & Thissen, 1996). If all of these conditions are met, testlets can be an excellent way to test some types of cognitive knowledge, but some care must be taken not to oversample areas of the content domain because several items are presented on the same topic. Two to three independent items per testlet set appears to maximize reliability (Norman, Bordage, Page, & Keane, 2006).

Other Selected-Response Formats

Extended-Matching

The extended-matching SR format extends the traditional matching format, making this item form useful to test higher-order knowledge (Case & Swanson, 1993). See Table 7.2 for an example. All matching items may be thought of as MCQs turned upside down, so that a common set of options is associated with a fixed set of items or questions. Each separate item of the EM set is scored as a free-standing item.

Like the traditional matching format, the extended-matching format is organized around a common set of options, all fitting the same general theme, and all providing plausible answers to a set of items designed to match this set of possible answers. See the NBME item writing guide (Case & Swanson, 1998) for good examples and discussion of this item type. As in traditional matching items, there should always be more options than items, so that a one-to-one correspondence is avoided. General directions for this form typically state: "Select each option once, more than once, or not at all."

Whereas traditional matching items generally test lower levels of the cognitive domain, like recall and recognition of facts, extended-matching items are ideal for testing higher-order cognitive knowledge relating to clinical situations such as clinical investigations, history taking, diagnoses, management, complications of therapy, outcomes of therapy, and so on. As a bonus, item writers, once they master the basics, may find EM items somewhat easier to create, since several related items are written around a common theme and at the same time. Also, EM items lend themselves to "mixing and matching" over different administrations of a test, since sometimes more item-option pairs than can be used on a single test are created for use on future tests.

For EM items of the clinical situation type, there must be a single common theme (e.g., diagnosis of related illnesses), with all the options fitting this common theme and all the items or questions relating to this theme, as in the example given in table 7.2. Most EM items briefly describe a clinical situation, presenting all the essentials facts of a patient problem or issue and a single focused question related to these clinical facts or findings. The items should be relatively short (no more than 2-4 sentences) and the options should be a short phrase or a single word.

The total number of options to use in EM sets is limited only by the constraints of answer sheet design (if machine-scored answer sheets are used). Many standard answer sheets are designed for a maximum of ten or fewer options, so the number of EM options has to be limited to a maximum number of options available on the answer sheet.

Some cautions are in order about EM items. Overuse of this item type on a single test could lead to an oversampling of some content areas to the detriment of other content areas. Since the EM format demands a common theme, it is likely that each EM item in the set will be classified as sampling content from the same general area. Many EM items on the same test could, therefore, oversample some content areas, while other important areas are overlooked (leading to the CU threat to validity).

True-False Formats

The true-false (TF) item format appears to be a simple SR format, requiring the examinee to answer either true or false to a simple proposition (e.g., Ebel & Frisbie, 1991). See Table 7.2 for an example of a true-false item. The TF item form requires an answer than can be absolutely defended as being more true than false or more false than true.

In health professions education, there are many examples of true-false items used to test very low level cognitive knowledge. In fact, many educators believe that the TF item form can be used to test only low-level cognitive knowledge (facts) and that most TF items test trivial content. While this may be an unfair criticism, there are many examples of TF items to support such a belief.

Measurement error due to random guessing on TF items is also a frequent criticism. If true-false items are well written and used in sufficient numbers on a test form (e.g. 50 or more items), measurement error due to blind guessing will be minimized. If these conditions are not met, random guessing may be a problem on TF tests. Like MCQs, TF items are best scored as “right or wrong,” with no formula scoring used to attempt to correct for guessing for most achievement testing settings in the health professions.

In fact, TF items can be used to test very high levels of cognitive knowledge (Ebel & Frisbie, 1991). The TF item has some strengths. For example, content-related validity evidence may be increased for TF items because many more TF items can be presented per hour of testing time compared to some other SR formats. Well written TF items can have sound psychometric properties, but TF items will almost always be less difficult than MCQs and the score reliability for TF items may be lower than for MCQs.

Creation of challenging, defensible TF items which measure higher-order knowledge is a challenging task. Some specialized skills pertain to TF item writing and these skills are rare.

Alternate-Choice Items (AC)

The Alternate-Choice (AC) item format (e.g., Downing, 1992) is a variant of the TF format. The AC form (see table 7.2 for example) requires less absoluteness of its truth or falsity and may, therefore, be more useful in classroom assessment in the health professions. However, the AC format is not used extensively, probably because it has many of the same limitations of the TF item form or at least is perceived to have these limitations.

Multiple True-False Items (MTF)

The Multiple True-False (MTF) item format looks like an MCQ but is scored like a series of TF items. See table 7.2 for an example. The MTF item consists of a stem, followed by several options, each of which must be answered true or false. Each item is scored as an independent item, as either correct or incorrect (Frisbie, 1992)

The strength of the MTF item is that it can test a number of propositions around a common theme (the stem) in an efficient manner. Some of the criticisms or perceived problems with TF items may apply to MTF items as well.

If MTF items are used together with other SR formats, such as MCQs or TF or EM item sets, it is important to consider how to fairly weight the MTF item scores relative to scores on other SR formats. The relative difference in time it takes to complete MTF items and MCQs is the issue. For example, if a test is composed of 40 MCQs and 4 MTF items each with 5 options (a total of 20 scorable units), what is the appropriate weight to assign these format scores when combining them into a single total test score? This weighting problem can be solved easily, but should be attended to, since -- in this example—the 40 MCQs are likely to take at least at least twice as long to answer as the 20 MTF scorable units.

Other SR Formats: Key Features

The SR formats discussed thus far in this chapter all aim to sample an achievement or ability construct comprehensively and representatively, such that valid inference can be made from item samples to population or domain knowledge. The Key Features (KF) format (Bordage & Page, 1987; Page, Bordage, & Allen, 1995) is a specialized written format which aims to test only the most critical or essential elements of decision-making about clinical cases. Thus, the purpose of key features-type assessment and the construct measured by KF cases differs considerably from typical achievement constructs. Farmer and Page (2005) present a practical overview of the principles associated with creating effective KF cases.

The KR format consists of a clinical vignette (1-3 paragraphs) describing a patient and all the clinical information needed to begin solving the patient's problem or problems. One or more CR and/or SR items follows this stimulus information; the examinee's task in these questions is to identify the most important or key elements associated with solving the patient's problem. The unique point of KF cases is that these items focus exclusively on only the most essential elements of problem solving, ignoring all other less essential elements. For example, KF items may ask the examinee to identify only the most critical working diagnoses, which laboratory investigations are most needed, and which one or more therapies is most or least helpful.

In some ways, the KF format is similar to the testlet format – a testlet with a unique purpose and form, that focus in on the most critical information or data needed (or not needed) to solve a clinical problem. But, there are major differences also. KF items usually allow for more than one correct answer, and they often mix CR with SR item forms. In this context, research suggests that 2-3 items per KF case maximizes reliability; use of fewer items per KF case reduces testing information and lowers reliability while using more than about 3 items per KF case provides only redundant information (Norman, Bordage, Page, & Keane, 2006). Like MCQ testlets, the case score (the sum of all individual item scores in each KF case) is the proper unit of analysis for KF cases.

Development of KF tests is challenging and labor-intensive with specialized training and experience needed for effective development. When the purpose of the assessment matches the considerable strengths of the KF format, the efforts needed to develop these specialized items is worthwhile.

SR Formats and Forms to Avoid

Some SR formats fail to perform well, despite the fact that they may have some intuitive appeal. Some SR forms have systematic and consistent problems, well documented in the research literature (e.g. Haladyna, Downing & Rodriguez, 2002), and should be avoided. See the NBME Item Writing Guide (Case & Swanson, 1998) for a good summary of item forms that are problematic and not recommended. Most of the problematic SR formats have the same psychometric issues: Items are either more difficult or less difficult and have lower item discrimination indices than comparable straightforward item forms. These problematic items also tend to be of lower reliability than comparable SR forms. But, these psychometric reasons may be secondary to the validity problems arising from use of item forms that may confuse or deliberately mislead examinees or provide cues to correct answers.

Complex Item Forms

One example is the complex MCQ format, sometimes called the K-type item format, following NBME convention (Case & Swanson, 1998). This is a familiar format in which the complex answer set consists of various combinations of single options. See Table 7.2 for an example. It was believed that this complex answer arrangement demanded use of complex or higher-order knowledge, but there is little or no evidence to support this belief.

In fact, this complex item form has some less than desirable psychometric properties and may also provide cues to the testwise examinee (e.g., Albanese, 1993; Haladyna, Downing, & Rodriguez, 2002). For example, once examinees learn how to take these items, they learn to eliminate some combined options readily because they know that one of the elements of the combination is false.

Variations of the complex format include the partial-K item which mixes some straightforward options and some complex options (Downing, 2005). Most testing organizations have eliminated these so-called complex formats from their tests.

Negative Items

Negation or the use of negative words is to be avoided in both item stems and item options. There are some legitimate uses of negative terms, such as the case of medications or procedures that are contraindicated; this use may be legitimate in that "contraindication" is a straightforward concept in health care domains.

Negative items tend to test trivial content at lower cognitive levels. One particularly bad form is to use a negative term in the stem of the item and also in one or more options – making the item nearly impossible to answer. While finding the negative instance is a time honored testing task, these items tend to be artificially more difficult than positively worded items testing the identical content and to discriminate less well – which lowers scale reliability (Haladyna, Downing, and Rodriguez, 2002).

Some item writers are tempted to take a textbook sentence or some phrase taken directly from a lecture or other instructional material, place a "not" or other negative term in the sentence, and then apply this negation to an item stem. For true-false items, this is a particular temptation, but one that should be avoided for all SR item forms.

Unfocused-Stem Items

MCQ stems of the type: "Which of the following statements are true?" are a time-honored tradition, especially in health professions education. Such open-ended, unfocused stems are not really questions at all. Rather, such MCQs tend to be multiple-true false items disguised as MCQs. In order to answer the item correctly, the examinee must first decide what question is actually being posed (if any), and then proceed to attempt to answer the question. Research shows that these types of open-ended, unfocused items do not work well (e.g., Downing, 2005), especially for less proficient examinees.

One helpful hint to item writers is that one should be able to answer the question even with all the options covered. Clearly, this is not possible for stems such as “Which of the following statements are true?”

Selected-Response Items: Summary Recommendations

Selected-response items are typified by multiple-choice items (MCQs) and true-false (TF) items. The best advise, based on a long research history, is to create straightforward positively worded SR items, with each item having a clearly stated testing point or objective; adhere to the standard principles of item writing. Complex or exotic formats should be avoided, since the complex form often interferes with measuring the content of interest. SR items should test at the cognitive level of instruction and be presented to examinees in sufficient numbers to adequately sample the achievement or ability domain. Three options is generally sufficient for MCQs, if the items are well targeted in difficulty and used in sufficient numbers on test forms. Random guessing is not usually a serious problem for well written SR tests. Right-wrong scoring is usually best. Attempts to correct raw scores for guessing with formula scores do not work well and may distort validity or bias scores by adding construct-irrelevant variance (CIV) to test scores, although in some cases formula scoring increases test scale reliability (e.g., Muijtjens, et al, 1999).

Summary and Conclusion

This chapter has overviewed some highlights of written testing in the health professions. Constructed-response and the selected-response item formats are used widely in health professions education for the assessment of cognitive achievement – primarily classroom-type achievement. Each format has strengths and limits, as summarized in this chapter.

Overall, the SR format – particularly its prototypic form, the MCQ – is most appropriate for nearly all achievement testing situations in health professions education. The SR form is extremely versatile in testing higher levels of cognitive knowledge, has a deep research base to support its validity, is efficient, and permits sound quality control measures. Effective MCQs can be securely stored for reuse. The principles used to create effective and defensible SR items are well established and there is a large research base to support validity for SR formats. SR can be administered in either paper-pencil formats or by computer.

CR items – particularly the short-answer essay – is appropriate for testing uncued written responses. Scoring for CR items is inherently subjective and procedures must be used to attempt to control essay rater biases. CR formats, such as short essay tests, may be appropriate for small classes of students, but scoring procedures must be carefully planned and executed in order to maximize score validity.

References

Albanese, M. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice*, 12(1), 28-33.

Baranowski, R.A. (2006). Item editing and item review. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of Test Development*, pp. 349-357. Mahwah, N.J.: Lawrence Erlbaum Associates.

Bordage, G., & Page, G. (1987). An alternative approach to PMPs: The key features concept. In I. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence* (pp. 57-75). Montreal, Canada: Heal.

Case, S., & Swanson, D. (1998). *Constructing written test questions for the basic and clinical sciences*. Philadelphia, PA: National Board of Medical Examiners.

Case, S. M., & Swanson, D. B. (1993). Extended matching items: A practical alternative to free response questions. *Teaching and Learning in Medicine, 5*(2), 107-115.

Downing, S.M. (2002a). Assessment of knowledge with written test forms. In Norman, G.R., Van der Vleuten, C.P.M., Newble, D.I. (Eds.). *International Handbook for Research in Medical Education* (pp. 647-672). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Downing, S. M. (2002b). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine, 77*(10), s103-104.

Downing, S.M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, 10: 133-143.

Downing, S.M. (2003). Guessing on selected-response examinations. *Medical Education*, 37, 670-671.

Downing, S.M. (2002c). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, 7, 235-241.

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

Downing, S.M. (2006). Twelve steps for effective test development. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of Test Development*, pp. 3-25. Mahwah, N.J.: Lawrence Erlbaum Associates.

Downing, S. M. & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*, 61-82

Downing, S.M., & Haladyna, T.M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education,* 38. 327-333.

Ebel, R.L. (1972). *Essentials of Educational Measurement*. Englewood Cliffs, NJ: Prentice Hall.

Ebel, R.L, & Frisbie, D.A. (1991). *Essentials of Educational Measurement*. Englewood Cliffs, NJ: Prentice Hall.

Farmer, E.A. & Page, G. (2005). A practical guide to assessing clinical decision-making skills using the key features approach. *Medical Education*, 39, 1188-1194.

Frisbie, D.A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice*, 5(4), 21-26.

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

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items* (3d Ed.).

Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance: A threat in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

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, 1*, 37–50.

Haladyna, T. M., & Downing, S. M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 51–78.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-334.

Linn, R.L. & Miller, M.D. (2005). *Measurement and Assessment in Teaching* (9th Ed.). Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp.

13-104). New York: American Council on Education and Macmillan.

Muijtjens, A.M.M., van Mameren, H., Hoogenboom, R.J.I.., Evers, J.L.H., & van der Vleuten, C.P.M. (1999). The effect of a 'don't know' option on test scores: number-right and formula scoring compared. *Medical Education*, 33: 267-275.

Norman, G., Bordage, G., Page, G., & Keane, D. (2006). How specific is case specificity? *Medical Education*. 40: 618-623.

Page, G., Bordage, G., & Allen, T. (1995). Developing key features problems and examinations to assess clinical decision making skills. *Academic Medicine,* 70, 194-201.

Rodriguez, M.C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24 (2): 3-13.

Thissen, D., & Wainer, H. (Eds.) (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practic*e, 15(1), 22-29.

Welch, C. (2006). Item/prompt development in performance testing. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of Test Development*, pp. 303-327. Mahwah, N.J.: Lawrence Erlbaum Associates.

Table 7.1

Constructed-Response and Selected-Response Item Formats: Strengths and Limitations

	Constructed-Response	**Selected-Response**
Strengths	• Non-cued writing • Easy to create • Logic, reasoning, steps in problem solving • Ease of partial credit scoring • In-depth assessment	• Broad representative content • Accurate, objective & reproducible scores • Defensibility • Accurate, timely feedback • Secure reuse of banked items • Efficient o Time o Cost o Information
Limitations	• Subjective human scoring • Limited breath of content • Reproducibility issues • Inefficient o Scoring time o Testing time	• Difficult to write well • Bad public relations o Guessing o Memorable

	o Information • Limited psychometrics and quality control	

Table 7.2

Examples of Constructed-response and Selected-response items

1. **Constructed-Response – Short Answer (Three sentences maximum)**

   1. Name and describe the function of each of the bones of the human inner ear.

2. **Constructed-Response – Long Answer (Five pages maximum)**

   2. Discuss the human inner ear, describing in detail how the inner ear structures relate to hearing.

3. **Selected-response – Traditional Multiple-Choice (MCQ)**

3. For a stable patient with a ventricular tachycardia of less than 150 beats per minute, which is the most appropriate first measure?

A. Intravenous lidocaine hydrochloride

B. Intravenous bretylium tosylate

C. Synchronized cardioversion

4. **Selected-response – True – False (TF)**

4. Random guessing is a major problem with the true – false testing format.

5. **Selected-response – Alternate-Choice (AC)**

5. If the number of items on a test is increased from 40 items to 60 items, the reliability of the 60 item test will most likely be:

a. Higher than the reliability of the 40 item test

b. Lower than the reliability of the 40 item test

6. **Selected-response – Multiple True-False (MTF)**

6. Which of the following increase the content-related validity evidence for an achievement test? (Mark each option as True or False)

a. Developing a detailed test blueprint

b. Training test item writers in item writing principles

c. Scoring the test using formulas that correct for guessing

d. Using some test items that are very hard and some items that are very easy

e. Selecting the most high discriminating previously used test items

7. Selected-response – Traditional Matching

7. Match each term on the left (A – C) with a definitions (A - D) on the right.

Each definition can used once, more than once, or not at all.

1. Hammer	A. Smallest bone in human body
2. Stirrup	B. Passes sound vibrations from eardrum
3. Pinna	C. Passes sound vibrations from malleus
	D. Visible part of outer ear
	E. Fluid-filled tubes attached to cochlea

8. Selected-Response – Extended Matching (EM)

8. Match each diagnosis (A-E) with the patient descriptions (1-3).

Each diagnosis can used once, more than once, or not at all.

A. Vasovagal reaction
B. Anaphylaxis
C. Lidocaine toxicity
D. Allergic contact dermatitis
E. Stroke

What is the most likely diagnosis for each patient who is undergoing or has recently undergone pure tumescent liposuction?

8_1. Immediately post op, a 49 year-old woman says that she “feels sick.” Her blood pressure is normal and her pulse is difficult to palpate; her skin is pale, cool and diaphoretic.

8_2. Six hours post op, a 25 year-old man is agitated and has tingling around his mouth. His speech is rapid.

8_3. During surgery, a 34 year-old woman says she "feels sick." She has generalized pruritus, her blood pressure begins to decrease and her pulse rate is rapid. Her skin is red and warm.

9. Selected-Response – Complex Multiple-Choice (Type-K)

9. What is the best treatment for a common cold (URI)?

Mark A if 1, 2, and 3 only are correct

Mark B if 1 & 3 only are correct

Mark C if 2 & 4 only are correct

Mark D if 4 only is correct

Mark E if all are correct

1. Rest
2. Fluids
3. Antihistamines
4. Decongestants

10. Selected-Response – Testlets or Context-Dependent Item Sets

10. One month after returning from Mexico, a 22 year-old college student presents with jaundice and abdominal pain.

10-_1. Which of the following will most likely develop in this patient?

A) fulminant hepatic failure
B) carrier state
C) chronic hepatitis
D) cirrhosis

10_2 What is the most likely route of transmission for this disease?

A) inhalation of contaminated air droplets

B) ingestion of contaminated food
C) mucosal exposure to bodily fluids
D) hematogenous spread

Table 7.3

A Revised Taxonomy of Multiple-Choice Item Writing Guidelines [1]

Content

1. Every item should reflect specific content and a single specific mental behavior, as called for in the test specifications.

2. Base each item on important content; avoid trivial content.

3. Use novel material to test higher level learning. Don’t use exact textbook language in test items, to avoid testing only recall of familiar words and phrases.

4. Keep the content of each item independent.

5. Avoid overspecific and over general content.

6. Avoid opinion-based items.

7. Avoid trick items.

8. Keep vocabulary simple and appropriate for the examinees tested.

Formatting Concerns

9. Use the question, completion, and best answer versions of conventional MC, the alternate choice, true-false, multiple true-false, matching, and the context-dependent item and item set formats, but avoid the complex MC format.

10. Format the item vertically, not horizontally.

Style Concerns

11. Edit and proof items.

12. Use correct grammar, punctuation, capitalization, and spelling.

13. Minimize the amount of reading in each item.

Table 7.3 (Cont)

A Revised Taxonomy of Multiple-Choice Item Writing Guidelines (Cont)

Stem

14. Ensure that the directions in the stem are very clear.

15. Include the central idea in the stem, not the options.

16. Avoid window dressing (excessive verbiage).

17. Word the stem positively, avoid negatives such as NOT or EXCEPT. If negative words are used, use the word cautiously and always ensure that the word appears capitalized and in bold type.

The Options

18. Develop as many effective choices as you can, but research suggests three is adequate.

19. Make sure that only of these choices is the right answer.

20. Vary the location of the right answer according to the number of choices.

Balance the answer key, insofar as possible, so that the correct answer appears an equal number of times in each answer position.

21. Place the choices in logical or numerical order.

22. Keep choices independent; choices should not be overlapping in meaning.

23. Keep choices homogeneous in content and grammatical structure.

24. Keep the length of choices about equal.

25. *None-of-the above* should be used carefully.

26. Avoid *All-of-the-above*.

Table 7.3 (Cont)

A Revised Taxonomy of Multiple-Choice Item Writing Guidelines (Cont)

27. Phrase choices positively; avoid negatives such as NOT.

28. Avoid giving clues to the right answer, such as:

    a. Specific determiners including *always, never, completely, and absolutely*.

    b. Clang associations, choices identical to or resembling words in the stem.

    c. Grammatical inconsistencies that cue the test-taker to the correct choice.

    d. Conspicuous correct choice.

    e. Pairs or triplets of options that clue the test-taker to the correct choice.

    f. Blatantly absurd, ridiculous options.

29. Make all distractors plausible.

30. Use typical errors of students to write your distractors.

31. Use humor if it is compatible with the teacher and the learning environment.

---

1. Quoted from and adapted from Haladyna, Downing, & Rodriquez, 2002, p. 312.

Table 7.4

Example of Analytic Scoring Rubric for Short-Answer Essay on Anatomy of Inner-Ear

Scale	Scale Point Description	Factual Accuracy	Structural Relationships	Writing
5	Excellent	All facts presented completely accurately	All structural relationships accurately described	Writing well organized, clear, grammatical
4	Good	Most facts correct	Most structural relationships correct	Writing fairly well organized, good clarity, mostly grammatical
3	Satisfactory	Many facts correct, some incorrect	Many structural relationships correct	Moderate organization and clarity, some grammatical errors
2	Marginal	Few facts correct	Few structural relationships correct	Little organization or clarity of writing, many grammatical errors
1	Unsatisfactory	No facts correct	No structural relationships correct	No organization or clarity, many serious grammatical errors