

Foundations of Education: an EMS Approach, 3rd Edition

Chapter 21 “Written Assessment”

Written by Mark Terry, MPA, NRP
Chief Certification Officer, National Registry of EMTs

For assistance in student evaluation, CoAEMSP has obtained permission from the National Association of EMS Educators (NAEMSE) to reproduce and make available to you Chapter 21 “Written Assessment” from the *Foundations of Education: An EMS Approach, 3rd Edition* book. The goal is to assist programs in meeting the **CAAHEP Standards:**



IV. Student and Graduate Evaluation/Assessment

A. Student Evaluation

1. Frequency and Purpose

Evaluation of students must be conducted on a recurrent basis and with sufficient frequency to provide both the students and program faculty with valid and timely indications of the students’ progress toward and achievement of the competencies and learning domains stated in the curriculum.

Achievement of the program competencies required for graduation must be assessed by criterion-referenced, summative, comprehensive final evaluations in all learning domains.

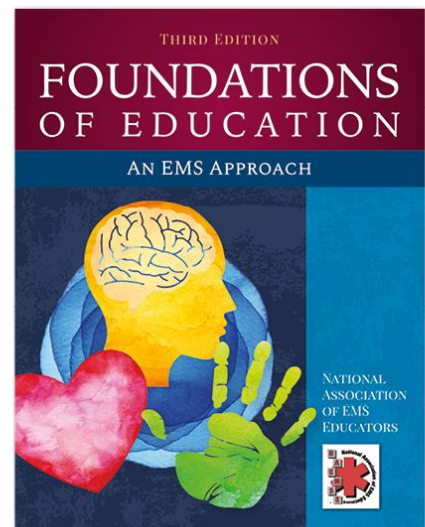
Part V of the book *Foundations of Education: An EMS Approach, 3rd Edition* is on student assessment, which includes:

PART V: STUDENT ASSESSMENT AND REMEDIATION.

- 20. Assessing Learning.
- 21. Written Assessment.
- 22. Other Assessment Tools.
- 23. Remediation.



To order a copy, or to learn more,
visit naemse.org or call 412-343-4775.



CHAPTER 21



Written Assessment

OBJECTIVES

At the conclusion of this chapter, the educator will be able to:

Cognitive Domain

1. Describe benefits and limitations of using written assessment tools in each domain of learning.
2. List steps to enhance reliability of written examinations.
3. Describe measures to improve written assessment validity.
4. Outline the steps to blueprint an examination.
5. Describe how to select appropriate items for an examination.
6. List effective test construction measures.
7. Distinguish between limited response and open (constructed) response items.
8. Explain the principles of constructing effective limited response items.
9. Given an example of a poorly selected (limited) response test item, edit it to improve its measurement precision.
10. Describe strategies to construct effective distractors for multiple choice questions.
11. Differentiate advantages and disadvantages of short-answer, essay, and fill-in-the-blank question types.
12. Describe advantages of formative assessments.
13. Outline effective test administration strategies.
14. Describe strategies to analyze examinations during a post-test review.
15. Distinguish between norm-referenced and criterion-referenced grading.
16. Describe methods to set a cut score for an examination.
17. Describe how item-response theory is used to establish passing criteria for computer adaptive testing.

Psychomotor Domain

There are no psychomotor objectives for this chapter.

Affective Domain

1. Defend the need to establish procedures that establish test validity and reliability.
2. Value the need to maintain test security.

“To those of you who received honors, awards, and distinctions, I say well done. And to the ‘C’ students, I say you too may one day become President of the United States.”

~ GEORGE W. BUSH

CHAPTER GOAL This chapter presents information on the construction, use, and analysis of written assessments.

Assessing students' knowledge is a key task for instructors. This is usually done through the use of written assignments and examinations. Each type of written assessment has its strengths, weaknesses, and implications for use.

The Written Assessment

One of the most common formal assessments of student performance is the written examination. The instructor can determine whether a written examination is the appropriate assessment instrument by considering the purpose of the assessment. Written examinations provide insight on student knowledge, but provide little information about a student's ability to perform a skill or consistently demonstrate a given attitude. Thus, written examinations are most useful for evaluating the cognitive domain. Written exams are not useful tools to evaluate psychomotor objectives. They can evaluate only lower levels within the affective domain. Grading, validating, or compiling results of written examinations is typically easier than other types of assessments. Because of this, written examinations are easily used with large numbers of students in a single class setting or across multiple classes. Written examinations that rely largely on multiple choice, true/false, and matching items are especially easy to grade; thus they are very useful with large classes. Students should also be exposed to testing strategies that mimic certification examinations to ensure they are prepared. As state and national examinations all have a multiple choice examination component (typically conducted using a computer system), the instructor should include that testing strategy in the emergency medical services (EMS) classroom—including the use of computer-based testing.

Appropriate selection of an assessment tool always depends on the proposed purpose of the assessment. Written examinations are best suited to answer questions such as the following:

- What does the student know about the subject?
- Which of the cognitive objectives has the student mastered?
- Does the student have the necessary knowledge to progress to more advanced material or complete the course of instruction?
- Have scheduled materials been presented adequately?

Properly constructed written exams can operate with high levels of reliability (an examination's ability to measure consistently). Because each student is being asked the same questions in the same way during the exam, consistent administration of the test is ensured. Most written examinations provide for consistent scoring, although there are challenges to ensuring grading reliability with some types of short-answer and essay questions. Because written examinations by nature are usually consistent in administration, reliability is mostly related to the quality of the individual items and scoring practices. This eases the processes of checking and monitoring reliability. Of course, poorly constructed items can, and usually do, have low reliability. Monitoring and improving reliability by evaluating and editing examination items promotes the appropriate function of these easy-to-use tools.

Similarly, high levels of validity (the ability of the exam to measure what it purports to measure) can be ensured by the use of carefully designed and written questions. Written exams generally encounter difficulties in this area. It is relatively easy to write examination items that assess low-level cognitive objectives, such as recall of key facts. Assessing higher-level thinking, such as problem solving or analysis, is more difficult with written examinations. Because of this, a common error for novice instructors is to assume that students have mastered higher-level objectives simply because they scored well on an examination filled with recall items. Efforts to ensure validity should include consideration of (1) the level of difficulty of required thinking (e.g., recall versus synthesis), (2) the breadth of the material covered (making sure the sampling of items is reasonable), and (3) the depth of the knowledge assessed by test items. The planning process to do this is referred to as **blueprinting**.

Each type of written examination item has its strengths, weaknesses, and implications for use. Proper use of written examinations requires an understanding of these strengths and weaknesses. Just as the selection of an assessment strategy is based on an understanding of the purpose of the assessment, the construction of a written examination requires the instructor to apply knowledge of test-item types to the objectives that the instructor is attempting to assess.

Construction of a Written Examination

Constructing well-written examination items from scratch is difficult, but can be learned and refined. Entry-level instructors should initially focus their efforts on using and improving existing examination items from their educational institutions and other

instructors. Textbook publishers and others are sources of exam items; however, many are low-level recall items and will need to be edited by the instructor. Using existing examination items still presents challenges for the instructor. Just because examination items are available does not mean that those items are valid or reliable, especially if they are used for several classes in a row. The instructor must always review and edit examination items for each class.

Some examination item banks and sources have been pretested for reliability. In these cases, the instructor can have more confidence in the items after reviewing the technical reports for item performance. The instructor should still exercise caution to ensure that the examination blueprint is a valid assessment for the material presented. The pilot population may be representative of the students being assessed, but those students may not be similar to students in any given instructor's class, which poses another potential problem. Still, pretested items are a valuable commodity to the instructor; it is much easier to begin with questions to modify than to create an entire exam from scratch.

Construction of a written examination consists of several key steps before the test can be put together. A flowchart of the examination construction process is shown in **FIGURE 21.1**. The first step is to carefully consider the purpose of the examination. The second step is to blueprint the examination, relating the breadth and depth of the examination to the stated objectives for the course. The third step is to develop or select draft examination items. Draft examination items are then reviewed by others and edited as needed (**FIGURE 21.2**). Reviewing exam items with other instructors or paramedics, and with the program medical director is important to verify the relationship of items to the objectives, to ensure their proper construction, to confirm the correct answer, and to discuss their relevance to practice. This review should be documented in some way to provide evidence that steps to ensure exam validity were taken.

For high-stakes exams, test items should be piloted to ensure items perform as expected.

Carefully Consider the Purpose of the Written Examination

The types of written examination items and the content of those items depend on the purpose of the assessment. Clear differences exist between the breadth of material used for formative exams and that used for summative exams. Elements to consider for purpose include the following:

- Are the subjects cumulative? In other words, should material from previous units be included? Including

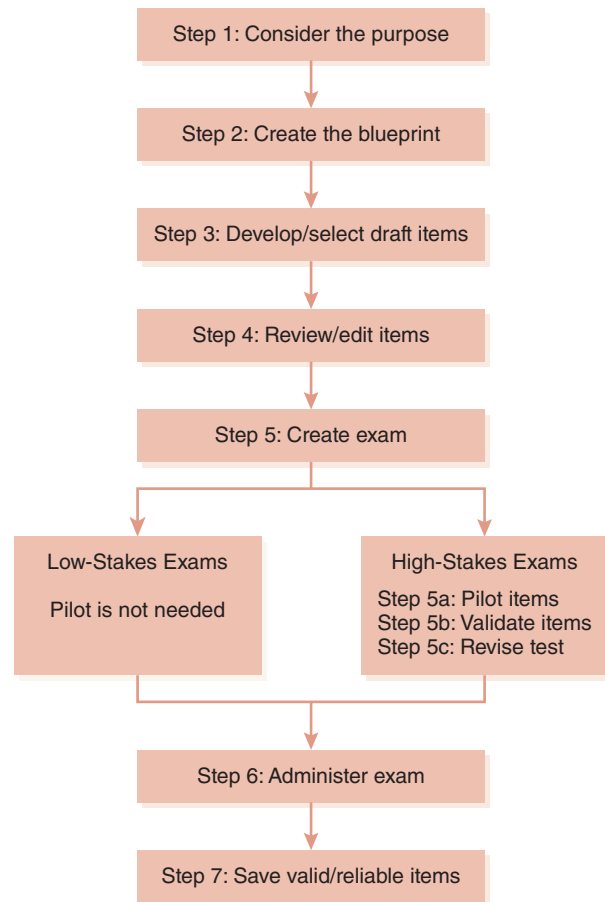


FIGURE 21.1 The examination construction process.

© Jones & Bartlett Learning



FIGURE 21.2 It is important that the instructor review exam items with knowledgeable sources (such as other instructors, paramedics, or the medical director) to verify that they are constructed properly and are relevant to the objectives and the practice.

© Nomad/Getty Images

some material from previous units is a good method to make sure students keep up with all material.

- What section(s) of the educational standards or course curriculum is being evaluated? To which objectives will the exam be tied? Ensuring the breadth

of the educational standards is well represented helps to prepare students for national certification examinations as well as for clinical practice.

- Are there limiting factors within the course design that affect assessment strategy? Examples might include available class time, need for immediate grading and feedback to students, or a large number of students with only one instructor to grade exams.

Blueprint the Examination

Once the purpose has been established, the next step is to blueprint the examination. The blueprinting process is conducted using the following steps:

1. The instructor lists the course objectives to be evaluated by the examination.
2. The instructor assigns a percentage of total questions (or points, if varying numbers of points are to be assigned to each question) written to cover each objective. If necessary, objectives can be grouped together and percentages assigned to each group.
3. The instructor selects the exam length. In general, use of more questions increases reliability, but very long tests (with more than approximately 150 multiple choice items) are much more difficult to develop and administer, and their use does not significantly improve reliability. Fatigue associated with very long examinations can reduce reliability and offset gains from the increased number of items. Conversely, examinations of less than 50 items frequently have difficulty proving reliability.
4. The instructor multiplies each section's percentage by the total exam length to determine the number of questions needed for each section.

Using the same strategy, the instructor should also construct a blueprint of the level of difficulty, based on the level of cognitive material that is being tested (i.e., recall versus synthesis). Instructors may wish to group objectives into categories so that the blueprint does not become overly complex. The instructor should balance the specificity of the blueprint with the complexity. It is rarely useful to specify item selection down to the individual item. If there are several categories in the blueprint with a single item, consolidation of the blueprint is usually indicated. For example, blueprints should indicate categories of items rather than specific items (for example, airway adjuncts rather than oral or nasal airway devices specifically).

Certification examinations usually publish the blueprint, also referred to as the test plan. Consulting

the blueprint for certification examinations can provide an idea of the level of detail required for an effective blueprint.

Generalizing

Instructors are sometimes tempted to assure themselves that students are prepared for examinations by emphasizing the content that is on the test. An example would be the instructor who knocks on the desk when a lecture point is covered on the exam. If students face items that are randomly selected from a large body of knowledge, it is more likely that one can generalize their knowledge of the entire body by their performance on the random sample. The ability to generalize is lost if the students know the sample ahead of time. For instance, if an instructor is trying to evaluate a body of 300 objectives, the instructor might select a random sample of 100 to include on the test. If the students do not know which 100 are on the test, the instructor can be assured that the students are preparing to address all 300. However, if the students know which 100 the instructor chose for the test, the instructor can only be assured that the students prepared for the 100 they knew would be on the test, not the wider body of knowledge. Test preparation should include all objectives in a module, not just those selected for the exam.

Develop or Select Draft Examination Items

From the blueprint, the process moves to the selection or drafting of items for use in the examination. The number of draft items collected should equal at least two times the number of items called for by the blueprint. Having more draft items than are called for in each area of the blueprint allows the editing process to select the most promising items to be refined. Some draft items will need extensive editing. If substantially more draft items are included than needed, items that require considerable rework can be eliminated if time becomes an issue.

Once an adequate number of draft items has been created or collected, the instructor can begin the review and editing stage. Only those items that have previously been validated can bypass this stage. The instructor should have colleagues and the medical director review the items and assist in the editing process.

Items taken from any commercially available test bank must also be reviewed and edited by the instructor before they are used. When possible, it is preferable for the instructor to employ unbiased editors who have

not drafted the selected examination items or presented the material to students. These editors should consider the following questions for each exam item:

- Are any grammatical or spelling corrections needed?
- Is the item clearly related to a stated course objective? A common mistake is to base items on instructors' presentation materials instead of on the course curriculum. One method to help counter this tendency is to have those providing draft selections also provide an annotated key that references each item to a course objective.
- Has the information/material been presented to the class in a lecture, reading assignment, or other means? Although it is appropriate to ask questions from reading assignments or nonclass content, care should be taken to ensure that the content is relevant to core objectives of the course. Some instructors also reference test items to a specific textbook reference to assist with later review and consideration. While this technique is useful to justify a correct answer, it is rarely helpful in justifying why a **distractor** (an incorrect answer option, also referred to as a foil) is incorrect. Additionally, higher-level cognitive or problem-solving items are rarely tied to specific reference in the text and may require more clinical judgment than is available in the text. Overreliance on textbook content can lead to an excessive number of recall-level items, particularly if textbook passages are used in the item.
- Is the item constructed appropriately? (See the following sections of this chapter on technical considerations for specific types of items.)
- What is the correct answer that is being sought? If it is a multiple choice item, is there only one correct (or clearly best) answer?

- Are the distractors clearly incorrect or substantially less correct than the key? The difficulty and reliability of an examination item frequently depends on the distractors, so these should receive the most attention.
- Are there any inadvertent hints to the correct answer?
- Is the level of difficulty of the question appropriate?

TEACHING TIP

Working cooperatively with other educators facilitates item development and editing. This could be as simple as a test-item exchange program between educators. A more complex approach would be for instructors to jointly host an item-writing workshop, inviting participation from a number of educational programs, and allowing all participants to use the results of a day's worth of item writing and editing.

Question Levels

Sometimes a simplified version of Bloom's taxonomy that includes three levels of test questions is used.

1. Recall questions assess understanding or memorization of facts.
2. Application questions require learners to categorize or apply their knowledge to new situations.
3. Problem-solving questions test the learner's ability to prioritize or make judgments using their knowledge of rules or principles in situations that vary from previously encountered situations.

CASE in Point

Blueprinting

An instructor is preparing a written examination to serve as a summative assessment of the cognitive material for a trauma unit that covers bleeding, soft-tissue trauma, burns, and chest trauma. The instructor prepares a blueprint of the exam.

The first step performed by the instructor is to gather information on the emphasis to be placed on each content area. The instructor begins by consulting the National Registry of EMTs' (NREMT) practice analysis, while using

three parameters (risk of harm, frequency, and difficulty). The instructor assesses each area for the risk of harm to the patient, assigning the greatest value to the riskiest, a lower value to the next riskiest, and so on. (Results are shown in **TABLE A**.) The instructor then does the same for frequency and perceived difficulty. Next, the instructor assesses the objectives of the particular course being evaluated. Counting objectives, the instructor notes that 30% of the module objectives are related to bleeding, 25% to soft-tissue trauma, 25% to burns, and 20% to chest trauma. The amount of

TABLE A NREMT Practice Analysis Example

Area	NREMT Practice Analysis			Curriculum Review		Expert Opinion		
	Risk of Harm	Frequency	Difficulty	Number of Objectives	Class Time Spent	Medical Director	Program Director	Adjusted Average
Bleeding	30	35	15	30	30	40	35	31
Soft-tissue trauma	25	50	10	25	30	30	35	29
Burns	20	10	35	25	20	10	15	19
Chest trauma	25	5	40	20	20	20	15	21
Total	100	100	100	100	100	100	100	100

class time spent on each content area is considered next, with the use of percentage allocation. The instructor also asks the medical director and the program coordinator to provide their opinions on the emphasis to be placed in each area, in terms of percentages. Results are averaged, and the totals are slightly adjusted by the instructor so the percentages add up to 100%. Table A shows the results.

The instructor next considers the level of thinking required for the objectives. The instructor assesses objectives written for the module and determines the percentage of objectives for each content area that is provided for each cognitive level. **TABLE B** shows the results.

The instructor next combines Tables A and B to determine the percentage of items that will be needed for each area and each level. This is calculated by multiplying the percentages assigned to each content area (Adjusted

Average column from Table A) and the percentage of each level shown on Table B. The results are shown in **TABLE C**.

The instructor had previously decided that the examination would consist of 100 items of equal weight (each item worth one point). The number of questions required is determined by multiplying the percentage in each column by 100 (the total number of items on the examination).

TABLE D shows the number of items needed for each level within each area.

The instructor now knows how many questions of each level and content area are needed for creation of a valid assessment of the student's knowledge of the content for this module of the course. The instructor can now select appropriate items from a test bank and proceed to the editing stage.

TABLE B Assessment of Course's Cognitive Objectives

Area	Remember (C1)	Apply (C3)	Evaluate (C5)
Bleeding	20%	30%	50%
Soft-tissue trauma	25%	30%	45%
Burns	25%	35%	40%
Chest trauma	25%	35%	40%

(continues)

CASE in Point (Continued)**TABLE C** Percentage of Items Needed for Each Content Area and Level

Area	Total	Remember (C1)	Apply (C3)	Evaluate (C5)
Bleeding	31%	6%	9%	16%
Soft-tissue trauma	29%	7%	9%	13%
Burns	19%	5%	7%	7%
Chest trauma	21%	5%	7%	9%

TABLE D Assignment of the Number of Problems for Each Level per Content Area

Area	Remember Items	Apply Items	Evaluate Items
Bleeding	6	9	16
Soft-tissue trauma	7	9	13
Burns	5	7	7
Chest trauma	5	7	9

Editing Multiple Choice Test Items

Many instructors find editing multiple choice test items particularly difficult. Collecting a group of instructors to jointly edit draft items can ease the task. This strategy is commonly used by large educational programs and those charged with certification examinations. Invited instructors are asked to bring a number of draft items as specified by the blueprint. The group then works together to edit the items, projecting the items so that all participants can see the editing process. Sharing editing tasks can improve the questions for a number of reasons:

- The bias and familiarity of the writer does not influence the revisions, as editing is shared between people who did not initially write the item. A single examination contributor introduces a significant challenge to reliability, as the interests and knowledge of that contributor becomes a major factor in the exam.
- Different options can be rapidly introduced and considered. Having multiple editors approach the task at the same time greatly reduces the cycle time of changes.
- Discussion of types of problems leads to more rapid solution when a number of items are edited together. The editing process speeds up over time.
- Frankly, more heads are better than one. The creativity of solutions builds as more editors are introduced.
- Local or regional bias or terminology is eliminated when individuals from a cross-section of the country work together on a national exam.
- Transparency of references and resources is ensured. For example, which textbooks or standards are appropriate to the exam?

Of course, there are limits to the benefits based on the number of editors and the time frame. Predetermined criteria for items can be established beforehand to clarify personal preferences and avoid arguments (such as those suggested in the following section). While a small group of editors is useful, a large group has difficulty reaching consensus. Fatigue limits creativity; so long editing sessions may be counterproductive. It is usually apparent when an editing team has “hit the wall” and fatigue sets into the group.

The Examination Construction Process

After items have been edited, the instructor can construct the examination. Instructors should consider the following guidelines regarding test construction:

- Be consistent in the use of punctuation and abbreviations. For example, periods are used at the end of the distractors if they complete a sentence, but not used if they are incomplete sentences.
- Use a consistent strategy to draw attention to material in the test (underline, bold, italics, or a combination).
- Use capital and lowercase formatting consistently for multiple choice items and for the first word of each option.
- If a separate answer sheet is to be used, ensure that the answer sheet and the test use consistent identification of options (e.g., 1, 2, 3, 4; A, B, C, D; or a, b, c, d).
- Provide clear and complete instructions for the examination—for example, whether the student can write on the test, whether there is a time limit, whether breaks are allowed, and (specifically for multiple choice items) whether there is only one correct answer versus whether students should select the *best* answer.
- For short-answer questions, students will commonly perceive the amount of space provided for the response as a suggestion for the length of the answer.
- The exam should be organized in a logical manner, with items from a similar content area grouped together. Some instructors believe that, similarly, the examination should begin with the easiest items, moving to harder items. The instructor should note that while these suggestions are intended to improve student satisfaction with the “flow” of the exam, certification examinations are often randomized. The use of greater randomization for summative examinations can help prepare students for certification examinations.
- If several items are related to a single scenario, then those items should follow a logical sequence. Care should be taken to ensure that a single incorrect answer does not jeopardize students’ ability to answer the next question correctly. In other words, although a single scenario can be used to set up a number of questions, each question should be capable of standing alone. Additionally, items linked to a single scenario should appear on the same page to avoid confusion.

Pilot use and validation should be conducted before an item is included in a high-stakes examination.¹ Items that demonstrate reliability and validity can be included in future exams, and items that fail can be returned to the editing process for improvement. A common mistake made by instructors is to pilot examination items using a single source that may not be representative of the intended audience. An example would be asking only other instructors their opinions on items for an entry-level examination. Although this may be useful to check content validity, other instructors are clearly not the same population that will be evaluated by the examination items. It is more useful in this situation to pilot the items using a population of other entry-level students.

For low- and moderate-stakes examinations, grading of the exams can be coupled with analysis. Two useful characteristics that can be identified in the analysis are difficulty level and item discrimination. **Difficulty level**, or difficulty index, is the percentage of students who answer each item correctly. **Item discrimination** is the degree to which a correct answer for a particular item is associated with high overall scores on the exam. Item discrimination is essentially a test of reliability. More on the analysis of written examinations is included later in this chapter.

One critical area for consideration with test items is the level of cognition and difficulty of the items that are used. Certification examinations, such as the NREMT, use items that test high-level problem solving. If lower levels such as recall and comprehension are the dominant form of test item within the educational program, then student performance on certification examinations will suffer. In identifying characteristics of educational programs that had high NREMT pass rates, Margolis noted three characteristics directly relating to written testing:

- Create and administer valid examinations that have been through a review process (such as qualitative analysis).
- Incorporate critical thinking and problem solving into all testing.
- Deploy predictive testing with analysis prior to certification.²

Items that test lower levels of cognitive objectives, such as recall, are relatively easy to construct. Therefore, there is a general tendency to choose items that evaluate lower levels of thinking than is intended by the writer. Instructors who are editing and reviewing items should be aware of this tendency and attempt to compensate by consistently ensuring that items evaluate problem solving and critical thinking.

Well-constructed and validated examination items are extremely valuable to the instructor and to the assessment process. This value is effectively destroyed if the security of items is compromised by the items being distributed to students in advance of the test. At the very least, such action converts an item that potentially evaluated high-level cognitive thinking into a simple memorization question. As such, validated items should be secured to the highest degree possible to preserve their usefulness.

Examination security can be breached in subtle ways. Letting students know which specific items are to be covered on a written examination is counterproductive in that students may then display false mastery of the material, which is not representative of their true abilities. A written assessment typically comprises a sample or “biopsy” of the objectives included in the course content. For this reason, if the student knows which specific knowledge areas are contained within the sample from which a broader conclusion is drawn, then the validity of the conclusion is challenged. In this case, the conclusion that the student has mastered the necessary material can extend only to what is directly assessed, and the conclusion that the student has mastered the broader areas from which the sample is drawn cannot be made. Although it is unavoidable that the instructor has previous knowledge of the test items, care must be exercised to not focus greater attention on specific content or items that will be covered in a future examination. An instructor does not *need* to know what specific items are covered on an examination, such as a licensure examination; the instructor needs to know only the objectives on which the examination is based. The idea of “teaching to the test” is often considered controversial. Instead of teaching directly to a test, both the teaching and test should be based on a common blueprint. When the examination and course are both derived from a common set of objectives, alignment is ensured.

Using Limited Response Items

The instructor may choose several different types of written examination items. Each offers its own advantages and disadvantages. Like other areas of evaluation of student performance, no single tool works for all situations. A combination of different types of examination items provides the strongest validity and reliability. Limited response (selected response) items contain a question or stem and require the student to select from answer options that are provided.

True/False Items

True/false items offer a complete statement with two possible choices: the statement is entirely true, or it is entirely false. True/false questions can present complex ideas to be evaluated, and they can be easily scored. Additionally, because students can complete them quickly, much more content can be tested in the allotted examination time with true/false questions

Examples of True/False Items of Various Cognitive Levels

Recall Item

T/F Positive-pressure ventilation is used for patients with inadequate spontaneous ventilation.

Note that this item is derived from a list of indications. Recall that “inadequate spontaneous ventilation” is a listed indication that enables the student to answer correctly.

Application Item

T/F A patient with cyanosis and a respiratory rate of 10 breaths per minute has adequate spontaneous ventilation.

Note that this item explores whether a situation fits within the category of adequate ventilation. The novelty of the description is important. If a study guide listed this situation as inadequate ventilation, the item would be testing recall. A higher level of cognition is tested by evaluating whether the student can correctly sort novel situations into the appropriate category.

Problem-Solving Item

T/F The head-tilt chin-lift maneuver is the preferred initial method of opening the airway for a child who is unresponsive and is not breathing after being struck in the head by a baseball.

Note that this item goes further than categorization. The student is given a novel situation and asked to evaluate a solution by applying several categories to the situation. First the student must categorize the situation into inadequate ventilation and recognize a need for spinal motion restriction. The student must then apply the indications and contraindications of the head-tilt chin-lift maneuver to the situation. Again, the novelty of the situation is important to preserve the assessment of higher levels of cognition. If a study guide said, “being struck in the head is a contraindication of the head-tilt chin-lift maneuver,” the item would test only recall.

Examples of How to Edit True/False Items

Poor

T/F Effective splinting always immobilizes the joints above and below the injury.

Better

T/F Effective splinting of long bone fractures immobilizes the joints above and below the injury.
(*Avoid absolutes.*)

Poor

T/F Oral airways are not used in responsive patients.

Better

T/F Oral airways are contraindicated in responsive patients.
(*Use positive statements to avoid confusion. Students taking a test will sometimes miss a single word in reading the item, and this presents a source for incorrect answers other than lack of knowledge.*)

than with other types of questions. One difficulty is that with only true or false as options, the statement must be either completely true or completely false. For example, if a statement is almost always true, the student is forced to guess whether the person writing the exam was thinking of the 99% of the time that the statement is true, or the 1% of the time that the statement is false. Another difficulty is that the chance of a random correct answer is 50%. In general, true/false questions tend to be very easy or very difficult. The result is that they do not always work well in discriminating between students of varying cognitive abilities. True/false items can be effectively combined with a short-answer format by asking students to justify their response. This can be used to assess higher levels of cognition and provide a framework that is slightly more directive than an open short answer.

True/false items should be written in the positive voice, avoiding negatively worded statements such as “is not.” It is also important to avoid absolute statements such as “always” or “never.” Very few absolute statements are entirely true, and students know this. The practice of taking statements directly out of the text should be avoided, as these are recall items of low difficulty. If a test is being taken by hand, to help eliminate problems in deciphering handwriting, instructors

should have students indicate true or false by circling or otherwise marking among provided selections, rather than having students write “T” or “F.”

Matching Items

Matching items typically present two columns of information with the intent that the test taker will select items from one column and match them to items in the second column to form correct statements or direct relationships. This strategy works best with terms and definitions, or with simple concepts and obvious relationships. However, this type of item can be confusing for the student unless clear instructions are provided.

This item does not work well when attempting to assess higher levels of cognitive learning, such as synthesis or evaluation.

Items to be matched should bear some similarity to each other to avoid making the correct response obvious. In other words, the list of responses should be homogeneous (e.g., do not mix doses with administration routes). With matching items, it is important for the instructor to provide clear instructions such as whether students will use each of the provided possible responses, whether one term can be used once or multiple times, or whether multiple answers are needed to complete a match. Poorly designed matching items are rather simple logic exercises, allowing students to use the process of elimination to greatly improve their chances of selecting the correct answer. The longer and more involved responses should be in the **stem** (the part of the item that is first offered, which may be written as a question or as an incomplete statement), keeping the responses short and simple. During construction, the instructor should take care to avoid giving grammatical cues to the correct answer. Matching sets should not exceed 15 items and should not break across pages. If the instructor is using scannable forms as answer sheets, the number of possible responses may be limited by the form used. This can be a significant limitation to the use of matching items.

Multiple Choice Items

Multiple choice items are commonly used in national and state certification examinations. Although multiple choice items are extremely easy to grade and demonstrate high interrater reliability (which is why these items are used for certification exams), they are difficult to properly construct. Multiple choice items consist of three main components: the stem, the distractors, and the key. The stem, as noted previously, is the part of the item that is first offered and can be written as a question or as an incomplete statement. The distractor is an incorrect answer designed to be a

Examples of How to Edit Matching Items

Poor

- | | |
|--------------------------|---|
| 1. Cyanosis | a. Used for unresponsive patients |
| 2. Nasal cannula | b. Used for airway control in responsive patients |
| 3. Oral airway | c. Delivers low-flow oxygen |
| 4. Bag-valve-mask | d. A sign of poor oxygenation |
| 5. Nasopharyngeal airway | e. Used to assist ventilation |

Better

- | | |
|--|-----------------------|
| 1. Provides high-flow supplemental oxygen | a. Bag-valve-mask |
| 2. Provides low-flow supplemental oxygen | b. Venturi mask |
| 3. Provides precise concentrations of oxygen | c. Nonrebreather mask |
| | d. Nasal cannula |

plausible alternative to the correct answer. The key is the correct (or best) answer to the stem.

Multiple choice items can be used to test both low and high levels of cognitive thinking, although constructing multiple choice items that evaluate high-level thinking is challenging. Multiple choice items are extremely easy to grade, and they allow for computer scoring of examinations. This makes it possible for a relatively large number of items to be used, thus increasing the reliability of the assessment instrument. On the other hand, because valid and reliable multiple choice items are difficult to

construct, the instructor is not able to rapidly develop these items. Constructing the examination items the night before the examination is simply not possible. Because a limited number of responses are allowed with multiple choice items, these items are unable to evaluate the thinking behind the selection of an answer. One variation on multiple choice items designed to overcome this limitation is to provide space within which the student can explain a selection, if the student believes that the provided information is not sufficient for a clear choice.

Suggested strategies for the proper construction of multiple choice items are as follows:

- Be on the watch for bias cueing (leading students to the correct answer by the way the stem is worded or from grammar choices).
- Avoid negatively worded stems. It is easy for students to misread negatively worded stems. Some educators propose that it is okay to use negatively worded stems when the concept tested is an important exception such as when *not* to do something. Medication contraindications are one example of this, such as “Nitroglycerin should *not* be administered to a patient with a systolic blood pressure of less than 90 mm Hg.” Where negative stems are needed, the negative word, such as “not” or “except,” should be italicized or boldfaced to draw attention.

In general, items should not build on previous items. Exceptions to this occur when the sequencing of steps is being assessed, or when a number of multiple choice items are related to a single, provided scenario. When a single scenario is used as the basis for several multiple choice items, the related items should be grouped together, should not break across pages, and may have a box drawn around the scenario and all related questions to ensure that students understand which questions belong to each scenario (FIGURE 21.3). Additional strategies include the following:

- Avoid questions written with a fill-in-the-blank segment in the middle of the stem; these are difficult to read.

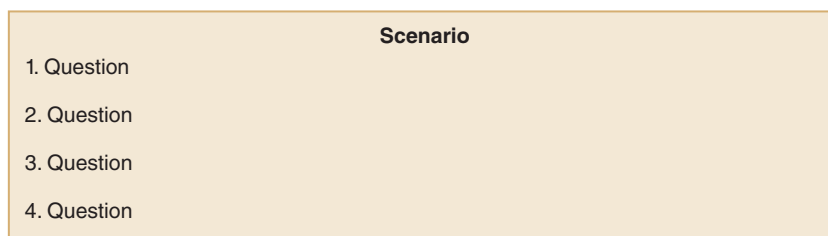


FIGURE 21.3 When a single scenario is used as the basis for several multiple choice items, related items should be grouped together, should not break across pages, and may have a box drawn around them, along with the scenario, to ensure that students understand which questions relate to each scenario.

- Avoid the use of “all of the above” or “none of the above” as an option. Recognition of one incorrect distractor immediately eliminates “all of the above” as the key. Recognition of more than one distractor as correct immediately indicates “all of the above” as the correct answer. Although “none of the above” presents less of a problem, it still presents the student with the ability to use simple logic instead of content knowledge to derive the correct answer. Often, rewriting the stem can prevent the use of “all of the above” and “none of the above” as choices. Additionally, if the instructions for the examination are to select the “best” answer, then use of “none of the above” is inappropriate, as one of the choices will be the best of those provided.
 - Avoid the use of “multiple multiple choice” items—questions that provide a list of possible components to the answer, with distractors and keys containing different combinations of components (for example, a question that includes answer options A–D, and then includes a list of options 1–4, along the lines of “1. A and B; 2. A, B, and C; etc.”). This type of item can be solved with basic knowledge, and these items are actually nothing more than a series of true/false questions. Instructors can easily convert “multiple multiple choice” items to a series of true/false items, thus correcting the deficiency.
 - Avoid overlapping responses. Overlapping responses present unnecessary difficulty to students.
- If the stem asks for a range and a distractor offers a single number, this can be immediately eliminated. Overlap of distractors into the correct range can be confusing for the student.
- Ensure that all distractors are approximately the same length. Common wisdom is that the longest option is usually the correct choice. This is because the writer of the item typically spends the most time with the wording of the correct option (to ensure that it is completely correct) and spends less time with the distractors.
 - Ensure that all distractors make grammatical sense. Frequent issues include problems with agreement of plural/singular and a/an. One way to easily avoid grammatical cueing is to use complete sentences as the stem.
 - Ensure that the correct answer is randomly distributed. A general tendency is for instructors to predominantly use (b) and (c) as the key.
 - Be aware that in constructing a multiple choice question, instructors tend to distribute the distractors so that two of the three are at the extreme positions, leaving the correct choice among the two middle answers (e.g., if the correct answer is 4, options would typically be 1, 3, 4, and 7). Students are aware of this tendency as well.
 - Place responses in a logical order. If the responses are assigned a numeric value, place the lowest numeric response as the first choice, the next highest as the second, and so forth.

Examples of Multiple Choice Items That Test Different Cognitive Levels

Recall

Which of the following parameters is included in the primary patient assessment?

- a. Blood pressure
- b. Level of consciousness
- c. Movement of distal extremities
- d. Bowel sounds

Application

Which of the following assessment findings is most helpful for determining the adequacy of ventilation?

- a. Skin color
- b. Heart rate
- c. Blood pressure
- d. Respiratory rate

Problem Solving

A patient from a motor vehicle collision presents with decreased level of consciousness, blood pressure of 170/100 mm Hg, heart rate of 60 beats per minute, and a respiratory rate of 10 breaths per minute. The skin is pale, cool, and moist. How should you administer oxygen to this patient?

- a. Bag-valve-mask
- b. Nasal cannula
- c. Nonrebreather face mask
- d. Venturi mask

Examples of Bias Cueing**Poor**

A patient presents as unresponsive, with no spontaneous respirations, after being hit in the head with a baseball bat. Which of the following would be the most appropriate device to use to secure the airway?

- a. Recovery position
- b. Oral airway
- c. Nasal airway
- d. Head-tilt chin-lift

(The term “device” in this example immediately eliminates choices a and d.)

Better

A patient presents as unresponsive, with no spontaneous respirations, after being hit in the head with a baseball bat. Which of the following would be the most appropriate means of securing the airway?

- a. Recovery position
- b. Oral airway
- c. Nasal airway
- d. Head-tilt chin-lift

(Bias cueing is removed by rewording the stem to remove the clue.)

Example of Multiple Choice with Fill-in-the-Blank**Poor**

You have initiated CPR on a patient in cardiac arrest. As soon as the equipment arrives, connecting ___ would be the next appropriate step.

- a. Oxygen
- b. Automatic external defibrillator
- c. Supraglottic airway
- d. Automatic transport ventilator

Better

You have initiated CPR on a patient in cardiac arrest. As soon as the equipment arrives, which of the following would be the next appropriate step?

- a. Oxygen
- b. Automatic external defibrillator
- c. Supraglottic airway
- d. Automatic transport ventilator

(The blank in the middle of the statement can present unnecessary confusion and is easily removed by rewording the stem.)

Example of Removing “All of the Above” as an Answer Choice**Poor**

Which of the following would be appropriate care for the patient with a serious chest injury from a motor vehicle collision?

- a. High-flow oxygen
- b. Spinal motion restriction
- c. Rapid transport
- d. All of the above

Better

Which of the following would **NOT** be appropriate care for the patient with a serious chest injury from a motor vehicle collision?

- a. High-flow oxygen
- b. Spinal motion restriction
- c. Rapid transport to the nearest trauma center
- d. Application of sandbags to the chest

(The easiest way to remove the “all of the above” option is to convert the stem into a negative phrase. In this case,

the negative “not” is in boldface and is capitalized to minimize confusion. Also, in the revised example, one of the distractors is lengthened, so the key is not the longest phrase among the choices.) The instructor should recognize that negative stems commonly have reliability problems as mistakes in reading produce measurable rates of error. Although it is easier to change to a negative stem, this may not be the best solution. Creative editing can correct this problem.

Better (without the Negative Stem)

Which of the following would be contraindicated in the patient with a serious chest injury from a motor vehicle collision?

- a. High-flow oxygen
- b. Spinal motion restriction
- c. Rapid transport to the nearest trauma center
- d. Application of sandbags to the chest

- Do not create words or abbreviations just to fill a response.
- Do not create humorous or ridiculous options just to fill space. The use of humor can create problems with reliability in addition to the fact that it may be perceived negatively by many students. As humor is culturally and often regionally based, the use of humor is a source of potential bias in the examination.
- All answers should be plausible to students. On later analysis, distractors that no students have selected

Example of Removing Multiple Multiples

Poor

Which of the following assessment findings is consistent with a patient who is suffering from hypoperfusion due to internal bleeding?

1. Warm and flushed skin
2. Rapid pulse rate
3. Low blood pressure
4. Anxiety
 - a. 1, 2, and 3
 - b. 1, 3, and 4
 - c. 1 and 3
 - d. 2, 3, and 4

Better

Which of the following assessment findings is **NOT** consistent with a patient who is suffering from hypoperfusion due to internal bleeding?

- a. Warm and flushed skin
- b. Rapid pulse rate
- c. Low blood pressure
- d. Anxiety

Another Option

Questions 12–15 refer to the following statement:
The following assessment findings are consistent with a patient who is suffering from hypoperfusion due to internal bleeding. Circle true or false for each assessment finding.

12.	Warm and flushed skin	True	False
13.	Rapid pulse rate	True	False
14.	Low blood pressure	True	False
15.	Anxiety	True	False

should be edited to improve the plausibility. Implausible distractors improve the odds of guessing the correct answer without the necessary knowledge.

Example of Fixing Overlapping Ranges

Poor

Which of the following is a normal respiratory rate for a patient who is 4 years old?

- a. 8–16
- b. 12–20
- c. 15–30
- d. 20–40

Better

Which of the following is a normal respiratory rate for a patient who is 4 years old?

- a. 10–15
- b. 16–30
- c. 31–50
- d. 80–100

(The overlapping ranges present an unnecessary difficulty. The situation is best avoided, even with the addition of a distractor that is far outside the range.)

Example of Ensuring Comparable Length of Answer Choices

Poor

A patient with severe respiratory distress should be transported in which of the following positions?

- a. Sitting, if the patient has a normal level of consciousness
- b. Supine
- c. Prone
- d. Recovery position

Better

A conscious patient with severe respiratory distress should be transported in which of the following positions?

- a. Sitting
- b. Supine
- c. Prone
- d. Recovery position

(Any necessary conditions for the key to be correct are moved to the stem, removing the obvious clue to the correct answer.)

Example of Removing Grammar Cues

Poor

A patient has an injury to the leg with severe pain and bone fragments protruding from the site of injury. This patient has an:

- a. closed fracture.
- b. open fracture.
- c. dislocation.
- d. sprain.

Better

A patient has an injury to the leg with severe pain and bone fragments protruding from the site of injury. This patient has a(n):

- a. closed fracture.
- b. open fracture.
- c. dislocation.
- d. sprain.

(Grammatical cueing, or grammar cueing, is easily avoided by using a complete sentence as the stem.)

Example of the Middle Value

Poor

Which of the following best expresses the range of respiratory rates considered normal for an infant?

- a. 12–20
- b. 15–30
- c. 25–50
- d. 50–70

Better

Which of the following best expresses the range of respiratory rates considered normal for an infant?

- a. 8–12
- b. 12–16
- c. 18–30
- d. 30–60

(Although all cases of the middle value being the correct choice do not need to be changed, the instructor should be aware of the tendency and take care to avoid patterns. Occasional use of an extreme value as the correct choice is appropriate. Although overlapping ranges are seen in this example, the student is being clearly asked to identify which range best describes normal, and the ranges in this case are taken directly from the 2011 AHA PALS provider manual: 12–16 normal for adolescents, 18–30 normal for school-age children, and 30–60 normal for infants.)

Challenges of Using the Full Sentence Stem

The examples in this chapter have used full sentence stems. While this practice easily eliminates grammar and bias cues when compared to blanks and incomplete sentences, use of full sentences as a stem also introduces challenges. Full sentences are longer than stems using incomplete sentences. The added length increases the time needed for examinations due to increased reading time. The added length also introduces a source of reliability problems from the unnecessary words. Writing stems as a full sentence is a reasonable practice for novice item writers, but experience with editing should enable more experienced writers to significantly shorten items through the use of incomplete sentences as stems. An example is provided:

A patient presents as unresponsive, with no spontaneous respirations, after being hit in the head with a baseball bat. Which of the following would be the most appropriate means of opening the airway? (33 words, 167 characters)

- a. Recovery position
- b. Oral airway
- c. Nasal airway
- d. Head-tilt chin-lift

(Editing to shorten the stem would produce a significantly shorter stem that is much easier to read and comprehend. Easy comprehension of key information in the stem is necessary for reliability.)

A patient struck on the head with a baseball bat presents as unresponsive and apneic. You should open the airway by using: (22 words, 100 characters)

- a. the recovery position.
- b. an oral airway.
- c. a nasal airway.
- d. the head-tilt chin-lift.

TEACHING TIP

One method of ensuring random distribution of the correct answer involves the use of a deck of cards. With each item, a card is selected:

- If the suit is hearts, A is used as the key.
- If the suit is clubs, B is used.
- If the suit is spades, C is used.
- If the suit is diamonds, D is used.

One must be sure to shuffle the deck before cards are chosen.

Using Open Response Items

Open-ended response items (constructed response items) can also offer advantages and disadvantages. In these question types the student must construct their own answer. A combination of different types of examination items provides the strongest validity and reliability.

Completion Items

Completion (also known as fill-in-the-blank) items are statements from which part of the information has been omitted; students must complete the statement. Enough information must be included for students to glean the intent of the statement without being led to the answer. One issue with open-response items arises when the meaning of the incomplete statement is unclear and several student responses emerge as correct, presenting a problem for the test grader. Items with unclear statements present a challenge for maintaining interrater reliability if more than one person is grading the exam. Completion items are not capable of evaluating higher-order thinking such as problem solving. These items are best used to evaluate recall, especially for key phrases that should be known verbatim or for definitions of key terms.

The provided answer space may present a problem for completion items. If one blank is used for each word of the correct response, the student is presented with a significant clue as to the answer. If only one blank is provided, students frequently assume that the answer consists of one word when multiple words are necessary. Either interpretation lowers reliability of the item.

Tips for writing completion items include the following:

- Omit significant words from the statement, but not so many that it is difficult for the student to determine the intent. For example, “An automated external defibrillator is used to treat ventricular ____.” is better than “A ____ is used to treat ____ fibrillation.” One method of ensuring this is to allow only one blank per completion item.
- Place the blank at the end of the sentence. This shortens the reading time and allows the student to derive the intent of the item before encountering the blank. For example, “____ is used to assess the percentage of hemoglobin that is oxygenated.” should be converted to “The percentage of hemoglobin that is oxygenated is assessed by ____.”
- As with other item types, avoid taking statements verbatim from textbooks or workbooks. It is

particularly tempting to construct completion items by copying a statement from the text and omitting key words. There are two problems with this practice. First, it ensures that the item evaluates only recall. Second, statements in texts are heavily dependent on context for the correct interpretation. Without that context, a single sentence frequently becomes ambiguous and difficult to complete the missing words.

- As with multiple choice items, be aware of possible cues from the grammar and sentence construction.
- Ambiguity of grading can be difficult with any open-response item. Because completion items require a short answer, it is difficult to evaluate the student’s thinking behind a particular response. This challenges the reliability of grading. For instance, consider the stem “Pulse oximetry is used to measure ____.” While the instructor may intend the answer to be “oxygen saturation,” reasonable responses may include circulation, shock, distal perfusion, oxygenation, and so on. It is difficult for an instructor to predetermine possible interpretations of an item without piloting the item or using reviewers.

Essay Items

Essay items pose a question or situation for which students are required to provide a relatively long, prose-style answer. Essay items are capable of assessing higher levels of cognitive thinking, but they also require that the student be capable of expressing this knowledge in coherent, written fashion. Essays can be used to effectively assess lower and middle levels within the affective domain. These questions also have the advantage of not being as easily susceptible to student guessing, although students may try to bluff. Because essay items are time consuming for students to complete and for instructors to grade, it is seldom practical to include more than a couple of essay items during a classroom assessment.

The sole use of essay questions on an examination presents a challenge to validity; this results from obvious problems with the breadth of material. Ensuring reliability during grading is difficult, as many factors other than knowledge can influence the assigned grade. In general, essay questions should be reserved for those objectives that cannot be effectively evaluated with limited response items.

The instructor should give their students advice for and practice with writing essays. This practice can be part of the formative assessment strategies. The instructor should not give students a choice of questions to answer during examinations. It will be difficult to

match the exam blueprint if different students answer different questions. Also, because some questions will be more difficult than others, the test could be unfair. When this choice is presented, each student is actually taking a different examination. Each essay question should be linked to a single objective; the student should avoid attempting to evaluate several objectives with one item.

Tips for writing essay items include the following:

- Avoid using essay items to evaluate recall of facts. Recall items, such as lists and definitions, are better evaluated using items that have fewer problems with grading reliability such as limited response items.
- Be clear in the task expected of students. For instance, “Discuss shock.” is much less clear than “Describe the various compensation mechanisms for shock.”
- In order to assess different levels of cognition, one useful strategy is to match the verbs used in the objective to the verbs used in the essay assignment.

Short-Answer Items

Between the essay question and the completion item lies the short-answer question. Short-answer questions are similar to essay questions, except that essay questions typically require multipage responses, and short-answer questions rarely exceed a full page. Depending on the stated objectives, it may also be desirable to avoid requiring the use of full sentences to respond to short-answer questions. Allowing students to use bulleted lists or outline forms may provide enough insight for the instructor to effectively assess knowledge, while not relying heavily on writing skills. Because they take less time and fewer writing skills for students to complete, more questions can be included. The strengths, weaknesses, and implications for short-answer questions are otherwise the same as for essay questions. In most cases, essay questions have little use in the EMS classroom compared with short-answer questions.

Writing essay items can be similar to writing short-answer items. It is important when writing a short-answer question to limit the scope of the question. When limiting the scope, the following may be helpful to convert essay items into short-answer items:

- Use a subset of clinical conditions; for example, “Describe compensatory mechanisms for *neurogenic* shock.”
- Describe the circumstances in more detail; for example, “The patient fell from a height of 20 feet. Describe the implications of selecting an appropriate destination for this patient.”

- Target the response by providing more detail about the expected response; for example, “Describe the pathophysiology of frostbite, paying particular attention to the role of vasoconstriction.”

TEACHING TIP

Both limited response items and short-answer items can assess higher levels within the cognitive domain. A simple rule regarding which type the instructor should choose is that short-answer or essay questions should be used when the time to prepare the examination is short and the time to grade the examination is long. When the time to prepare the examination is long and the time to grade the examination is short, limited response items (such as multiple choice) are the preferred tool.

Alternate Item Formats

Alternate item format may be seen on some exams. These question types are frequently used to assess learners’ ability to interpret information at higher levels. Alternate item formats include multiple response, drag-and-drop (ordered response), and media-enhanced items. Media-enhanced items, including hot-spot, chart/exhibit, audio item format, and graphic distractor options are variations on the traditional multiple choice question (**TABLE 21.1**).^{3,4} These question types are particularly difficult if the students have not previously encountered them.

Homework and Research Projects

A variety of homework and research projects can also add to formative and summative student assessment. Some options are discussed here.

Homework

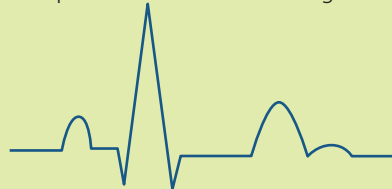
One tool that can be useful as a formative assessment is the routine assignment of homework to be completed by students. Homework should be spread out relatively evenly across the course. Each assignment need not be graded, but many students will interpret the lack of grade impact as lack of importance. Assigning a nominal grade impact to a random selection of homework assignments can counter this tendency. The instructor should review homework assignments for level of difficulty and should include a mix of easy and difficult items. Encouraging students to collaborate on homework can be a useful practice that helps





to build teamwork and peer learning. Frequent assignments, which help to build regular study habits in students, provide the instructor with regular, formative feedback on student progress.

Examples of homework assignments for the EMS classroom include assignments from workbooks, completion and definition worksheets, short research projects, case studies, concept maps, targeted

discussion in an online forum (discussion board or blog), writing a summary of key points from lecture, and description of care for a supplied scenario. Electronic homework assignments can include questions embedded in narrated lectures, online quizzes, and virtual simulations. Homework assignments also provide students with examples of what types of problems they will be expected to solve for summative

TABLE 21.1 Alternate Item Types

Alternate Item Type	Description	Example								
Multiple response items	Student must select all of the options that are correct. It is possible to have a single correct response, more than one correct response, or all responses are correct. Note that these items allow varied grading strategies for partial credit, which the instructor should make clear in the exam instructions.	<p>A 72-year-old female has abdominal pain after a motor vehicle crash. Which signs and symptoms would you anticipate if she is developing shock?</p> <p>Select all that apply:</p> <ul style="list-style-type: none"> <input type="checkbox"/> Anxious appearance <input type="checkbox"/> Increased respirations <input type="checkbox"/> Pale skin color <input type="checkbox"/> Slowing heart rate 								
Hot-spot items	Test taker selects a specific area on a figure, graph, or diagram to illustrate the correct answer. Note that these items may also allow multiple responses.	<p>Select the J point on the electrocardiogram below.</p>  <p style="text-align: right; font-size: small;">© Jones & Bartlett Learning</p>								
Drag-and-drop (ordered response)	Candidate selects and moves items into a specified order or sequence.	<p>A 22-year-old male with an apparent opiate overdose is unresponsive, with a respiratory rate of 6 breaths per minute. Arrange the following steps in the order the EMT should perform them.</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="background-color: #0070C0; color: white;">Unordered Options</th> <th style="background-color: #0070C0; color: white;">Ordered Response (Answer)</th> </tr> </thead> <tbody> <tr> <td>Administer intranasal naloxone</td> <td>Insert an oropharyngeal airway</td> </tr> <tr> <td>Begin bag-mask ventilation</td> <td>Begin bag-mask ventilation</td> </tr> <tr> <td>Insert an oropharyngeal airway</td> <td>Administer intranasal naloxone</td> </tr> </tbody> </table>	Unordered Options	Ordered Response (Answer)	Administer intranasal naloxone	Insert an oropharyngeal airway	Begin bag-mask ventilation	Begin bag-mask ventilation	Insert an oropharyngeal airway	Administer intranasal naloxone
Unordered Options	Ordered Response (Answer)									
Administer intranasal naloxone	Insert an oropharyngeal airway									
Begin bag-mask ventilation	Begin bag-mask ventilation									
Insert an oropharyngeal airway	Administer intranasal naloxone									
Audio item	Test taker listens to a sound clip using headphones and selects the correct option.	<p>A 65-year-old female complains of sudden onset difficulty breathing. Her breath sounds are as follows (audio clip of crackles is played). Which intervention is indicated?</p> <ul style="list-style-type: none"> a. Albuterol updraft b. Epinephrine IM c. Magnesium sulfate IV d. Nitroglycerin SL 								

Alternate Item Type	Description	Example								
Graphic option items	Item distractors are images rather than words.	<p>A 65-year-old has fever and tachypnea. Which rash would indicate the need for rescuers to don an N-95 mask?</p> <p>a.  © Joel zatz/Alamy Stock Photo</p> <p>b.  © Mediscan/Alamy Stock Photo</p> <p>c.  © Allan Harris/Medical Images</p> <p>d.  © LeventKonuk/iStockphoto</p>								
Chart/exhibit	The candidate interprets information within a chart or exhibit to solve a problem.	<p>The paramedic receives a patient from an urgent care center who complains of weakness, weight gain, and strong-smelling urine. The lab report shows serum:</p> <table border="1" data-bbox="885 1354 1421 1549"> <tbody> <tr> <td>Sodium</td> <td>145 mEq/L</td> </tr> <tr> <td>Potassium</td> <td>6.2 mEq/L</td> </tr> <tr> <td>Hemoglobin</td> <td>14 g/dL</td> </tr> <tr> <td>Leukocytes</td> <td>10,000/mm³</td> </tr> </tbody> </table> <p>Which should the paramedic assess first?</p> <p>a. Capnography b. Electrocardiogram c. Pupil response d. Temperature</p>	Sodium	145 mEq/L	Potassium	6.2 mEq/L	Hemoglobin	14 g/dL	Leukocytes	10,000/mm ³
Sodium	145 mEq/L									
Potassium	6.2 mEq/L									
Hemoglobin	14 g/dL									
Leukocytes	10,000/mm ³									
Video items	A question is asked based on a video clip of a situation or procedure that is played.	<p>A video illustrating defibrillation with an automated external defibrillator (AED) is played and the candidate is asked:</p> <p>Which action in this sequence was incorrect?</p> <p>a. CPR continued during analysis. b. Pads applied before AED turned on. c. Pulse was checked after defibrillation. d. Rescuer touched patient as shock delivered.</p>								

assessments. To be effective as formative assessments, homework assignments must be graded and returned to students in a timely manner.

Research Project Assignments

Project assignments based on students' own research are another means of assessing the ability of students to synthesize information. These assignments are a tool for assessing higher-level cognitive learning. Individual projects allow students to use their own specific learning preferences to complete the assignment. Research projects promote student autonomy, enhance student confidence, and encourage independent learning.

Assignment or choice of topic is an important, yet commonly overlooked, component of the project assignment. Allowing students to choose a topic that appeals to them is appropriate, but the instructor must be an active part of the topic selection and determination of project scope. Students should not waste valuable time on consideration of topics. One way of avoiding this scenario is to prepare a list of potential topics from which students can choose. Controlling the project scope is necessary to ensure that projects assigned to different students are roughly equivalent in terms of difficulty. Many instructors require that students obtain approval on project scope early in the process.

Another option is to prepare a rubric that clearly outlines guidelines for determination of grades based on the amount of work that students complete. For example, "To earn a C, the student will complete a written paper of at least 10 pages and a classroom presentation; to earn a B, the student will also complete at least one optional activity; and to earn an A, the student will complete at least two additional optional activities. Optional activities include reporting on an interview of a local medical director, creating a project-related website, completing a survey of at least 20 local EMS providers, and creating and demonstrating a working mechanical model related to a particular topic."

A measure of negotiation between the instructor and the student is appropriate in determining scope while still allowing students to express their own talents and learning preferences. It is also helpful for the instructor to reinforce relevance by creating realistic writing scenarios, such as, "Your medical director has asked you to submit a new protocol for the treatment of anaphylaxis. Please submit your protocol, which should include both assessment and treatment sections. Provide at least five sources from peer-reviewed medical journals that support the care you propose." One problem with project assignments is the tendency of instructors to base the grade on product rather than on process. In most cases, the process used by students

to prepare a project is just as important as the product itself. One way that instructors can avoid this trap is by requiring students to submit intermediate steps for consideration and possible impact on grade. An example is to have a check-in for the following steps: (1) description of title, purpose, and major points; (2) sources, data, and references; (3) outline; (4) first draft; and (5) final version.

The process of grading project assignments is essentially the same as that used to grade essay items on written exams. Recommendations provided in the section *Grading Essays* later in this chapter can be applied to written components of the project. Criteria for grading projects should be clearly communicated to students, as noted in the section, *Grading Strategies*. The grade may contain components that measure the quality of the product, as well as the effectiveness of the process.

Project assignments are best viewed as a combination of a learning tool and an assessment tool. As a learning tool, project assignments result in learning that is customized to the individual student's talents and preferences. Project assignments emphasize critical thinking, independent learning, and use of research skills. As an assessment tool, project assignments permit assessment of high-level cognitive objectives and, in some cases, affective objectives.

Administering Written Examinations

The administration of written exams requires careful attention. An inappropriate environment or ineffective method of administration can significantly impact exam validity.

Environmental Considerations

The classroom set-up for a written examination is essentially the same as that used for a traditional lecture format, with the students seated in rows (**FIGURE 21.4**). Students should be seated far enough apart to discourage them from looking at each others' papers. Exam proctors should walk the room from time to time so that they are able to see students' faces as well as observe students' space and activity. Appropriate temperature and lighting should be ensured. Special attention should be given to providing a quiet environment.

For examinations that last longer than an hour, the instructor should set clear rules for restroom breaks. It is helpful to have extra copies of the examination and answer sheets, scratch paper, and pencils readily available.



FIGURE 21.4 Answer sheets and test booklets (facedown) can be placed at student seats before the examination. Student notebooks and backpacks should be placed at the back or side of the room.

Courtesy of St. Charles County Ambulance District.

Proctoring

An instructor should supervise written examinations to discourage cheating and to address problems or process questions as they arise. Lead instructors communicate the importance of examinations by **proctoring** the examination themselves. Proctors should arrive early and should be prepared to leave late. During the examination, the proctor should monitor the room without hovering over students. The proctor should have a strategy for addressing questions asked by students during the exam. One common strategy is to allow the proctor to answer only questions regarding examination process—not questions related to examination content. Proctors should be cognizant that any communication of content to a student who asks a question gives an advantage to that student over those who did not ask or attempt to fish for clues. The proctor should keep students apprised as to the time by having a clock in the room, writing (and updating) the time on a whiteboard, or periodically announcing the time remaining for the test.

Appropriate proctoring helps to ensure the integrity of the examination. Cheating can flourish in an unsupervised environment. The proctor must maintain security of testing materials. The ability to look at other students' answers can be prevented by proper seating arrangements. Cell phones, smartwatches, and other recording devices should be prohibited in the testing environment. Any notes or calculations during the examination should not leave the testing environment. Silence during examinations discourages covert communications. Use of multiple versions of an exam, with differing arrangements of item sequence and

distractors, discourages organized efforts by groups of students to each memorize parts of an examination and later reconstruct the examination. Another means to defeat that form of cheating is to revise examinations after each administration. Test development software can make creation of multiple randomized versions easier. Diligent observation by the instructor, combined with clear expectations of integrity, are key to preventing cheating.

Computer-Based Testing

Technological developments have enabled more widespread use of computer-based testing in educational settings. Once reserved for high-stakes examinations, these techniques are increasingly available for classroom instructors to incorporate. Testing centers in community colleges and other environments can offer secured, computer-based testing environments, which may be appropriate for high-stakes summative examinations. Other variations exist for use within the classroom. Use of a learning management system (LMS), such as Moodle, Canvas, or Blackboard, may include testing modules. This allows for distributed methods of formative assessments and quizzes. Within the classroom, small, remote, polling devices or audience response systems can connect to a presentation system. This system can allow an additional layer of interaction within the classroom discussion that can blend a tracked, formative assessment with an informal discussion.

A major advantage of computer-based testing systems is that these systems remove a potential source for error in grading and item analysis. By direct entry into the system, grading and analysis are more efficient. With varying degrees of security, tests can be offered in multiple locations—even in the student's home at a convenient time. Computer systems also allow a greater variety of media to be attached to test items such as pictures, audio, and video. Different versions of the test can be offered to students, increasing security. Using different versions of the examination is enabled by the substantial increases in efficiency of grading. Feedback can be offered instantaneously, which is particularly valuable for formative assessments.⁵

These systems typically favor limited response items. This limits the range of items available to instructors. If large numbers of versions are used, an extensive item bank may be required. Item security may be difficult to maintain, particularly in formative exams linked to an LMS. Cheating may be difficult to monitor if the examination is delivered in a distributed mode, although several varieties of exam security and verification of identity may be used in formal testing

centers. Some vendors offer examinations that include online remote proctoring in which the examination proctors monitor the student remotely through the use of technology, such as webcams. The examination process that uses computer-based testing can be subject to a variety of technical difficulties that are not present in paper-and-pencil versions, such as network outages.

Some self-directed educational programs build computer-based assessment directly into the learning algorithms. Assessments and performance in electronic simulations guide content. These programs, such as the American Heart Association Heartcode Advanced Cardiovascular Life Support (ACLS) and Pediatric Advanced Life Support (PALS), combine assessment and learning activities into computer-based simulations. This can effectively combine formative assessment, learning, and summative assessment into a blended set of activities that are seamless to the learner and yet contain sophisticated recorded evaluations of learner abilities.

The NREMT and other certification bodies use computer testing systems to deliver certification examinations. Although most computer-based exams use traditional testing theory and are linear, NREMT actually uses a different testing format called **item-response theory (IRT)**. This exam format allows more precise measurement with fewer items. More information on IRT is presented later in this chapter. Although IRT is not usually an option for use in EMS classrooms, it seems reasonable that instructors preparing students for NREMT certification would build a degree of computer-based assessment into their programs. The rapid pace of change in technological environments ensures that developments in technology-assisted assessments will outpace the ability of any text to adequately describe current capabilities.

Time Limits

Each student will take a different amount of time to complete the examination. To exert some measure of control over the time spent on the examination, the instructor must set some limits on time. Setting time limits for examinations is a legitimate strategy for (1) preparing students for certifying examinations and (2) evaluating students' ability to think quickly. The drawback to setting time limits is that some students struggle to complete the examination in the time allowed. In estimating the amount of time a student is given to complete an examination, the instructor can give students four times as long as it takes the instructor to complete the test. As an alternative, Barbara Gross Davis, in *Tools for Teaching*,⁶ suggests the following timing strategies:

- Allow half a minute per true/false item.
- Allow 1 minute per multiple choice item.
- Allow 2 minutes per short-answer item.
- Allow 10 to 15 minutes per limited essay item.
- Allow 30 minutes per broader essay item.
- Allow 5 to 10 minutes for students to review their work.
- Factor in time to distribute and collect tests.

It is critical to note that the examination should be designed and administered to assess only those abilities necessary and not inadvertently depend on unrelated abilities. This is particularly important when considering learning disabilities. Examinations can be unreasonably dependent on reading abilities. One strategy to accommodate documented learning disabilities would be to extend the time allowed for an examination. In some cases, a reader would be appropriate for a written examination. The evaluation of disabilities and compliance with the Americans with Disabilities Act (ADA) is beyond the scope of this text. Expert evaluation of the situation may be necessary, in which case the instructor should consult with available experts in the educational setting.

Analysis of Written Examinations

The analysis and potential revision of written exams are important steps in improving student assessment. The type of evaluation depends on the examination stakes as well as available resources.

Post-Test Review

A useful strategy after an examination has been administered is to allow class time for students to review the examination as a group with the instructor. This review highlights areas of weakness for individual students, as well as for the class as a whole. Review can also help the instructor to identify areas where the presentation of material did not adequately prepare students for mastery of the stated objectives. It can serve to alleviate concerns about bias when students see what items other students missed. A climate of fairness is promoted when students can discuss questions, answers, or the wording of a question. Although some instructors allow students to retain the examination after classroom discussion, this practice greatly reduces the validity of test items that are reused. Even on low-stakes examinations, students who have access to the previous classes' exams can develop a false sense

of security, thinking they are familiar with the content when in fact they are only recognizing items they have seen on previous exams. Teachers may be misled about the students' understanding of material based on answers obtained on previous exams. Additionally, it is neither time nor cost effective to develop new exams for each class, even for low-stakes exams. Conducting a classroom discussion breaches examination security, but the breach is less significant than when students are allowed to retain copies of the examination.

Pilot use and previous validation may not be possible for all examinations, but they should be conducted before an item is included in a high-stakes examination. One possible strategy for pilot use is to present pilot items for formative assessments, such as quizzes. Another is to have an examination include several (generally not more than 10%) pilot items that do not count toward the exam score. These should be interspersed among regular items. Pilot items that demonstrate reliability and validity can then be included in future examinations. Pilot items that fail validation can be returned to the editing process for revision, guided by pilot data.

Difficulty Level and Discrimination Index

For low- or moderate-stakes examinations, grading of the examination is coupled with validation of test items. Validation is particularly applicable to limited response items such as true/false, multiple choice, and matching questions. As mentioned earlier, the two characteristics of tests that are useful in validation are difficulty level and item discrimination index (also called the discrimination ratio). (Recall that the difficulty level is the percentage of students who answer each item correctly; the item discrimination index compares the performance of those who scored well on the exam with the performance of those who did not score well on each exam item.) Computerized programs will perform the necessary calculations (**FIGURE 21.5**), but the same measurements can be easily calculated manually, which means the instructor can validate items even when not administering examinations by computer.

Students may find an item difficult for numerous reasons, including that the item may be poorly worded. Adding the discrimination index into the analysis for potential revision of items separates those items that have questionable reliability and validity from those that are appropriately constructed, yet challenging. Items that have extreme difficulty levels (either high or low) will not discriminate as well as those with a difficulty level near 50%. As a result, different thresholds are used to indicate the need for revision depending

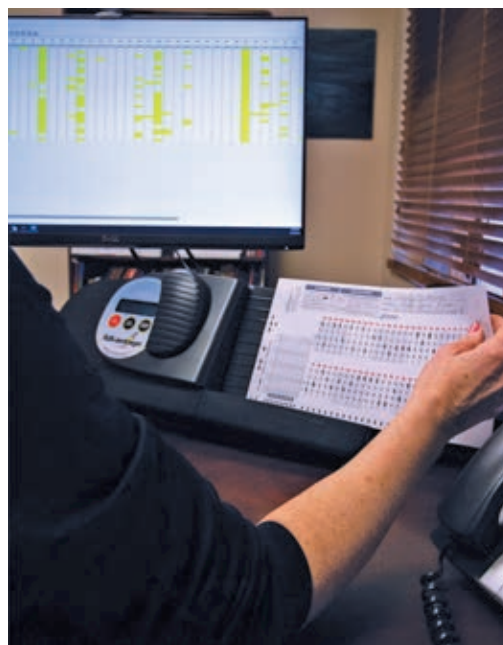


FIGURE 21.5 Scannable answer sheets and grading software can facilitate quick scoring of multiple choice exams and provide a means of performing item analysis.

Courtesy of St. Charles County Ambulance District.

on the difficulty level of the item. When the difficulty level and the discrimination index are used, the process shown in **FIGURE 21.6** can help identify items that need revision.⁶

Negative discrimination indices indicate that students who scored well overall did worse on those

Calculating the Difficulty Level

The following procedure can be used to calculate the difficulty level and the discrimination index for limited response items, such as multiple choice or true/false questions.

To calculate the item difficulty, the instructor should calculate the percentage of students who had correct responses. The formula for this is:

$$ID = (C/T) \times 100$$

where ID is the item difficulty, C is the number of correct responses, and T is the total number of students who took the examination.

For example, if 30 students took the exam and 20 answered the item correctly, the difficulty level/index would be 67%. The goal is to use only a few items that more than 90% or less than 30% of students answer correctly.⁶

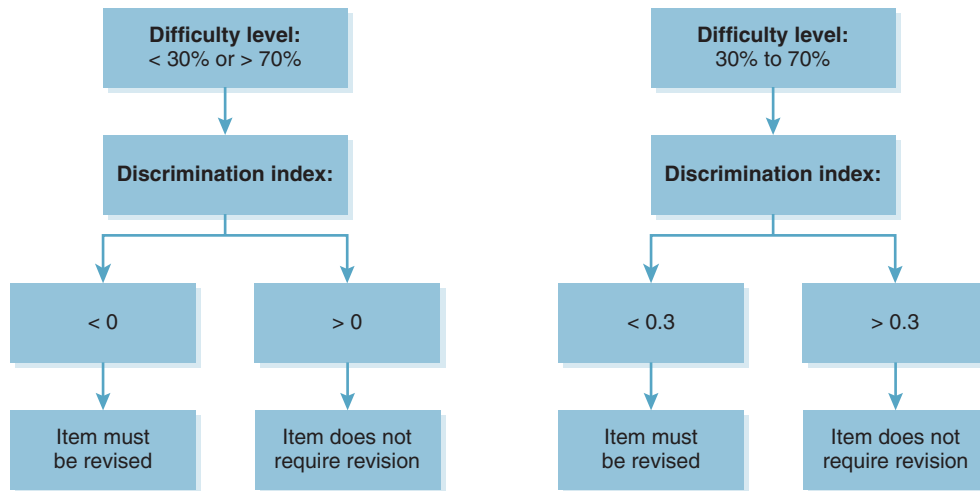


FIGURE 21.6 The difficulty level and the discrimination index can help the instructor identify test items that need revision.

Data from Davis, Barbara G. 2001. *Tools for Teaching*. San Francisco: Jossey-Bass.

particular questions than did students who did not score well overall. Items with a negative discrimination index should be reviewed for validity and revised before they are used again. A common cause of a strongly negative item discrimination index (near to -1.0) is an incorrectly

keyed item. Editing of examination items should extend to a check of the answer key as well.

When analyzing items, there are several components of the item that should be considered. Problems with the stem are the most obvious and common to

Calculating the Item Discrimination

The purpose of calculating the item discrimination is to compare the response of the high exam performers on each item to the response of the low exam performers. If an item is constructed correctly, the instructor can expect the high scorers to get the item correct and the low scorers to miss the item. This would be referred to as a *positive discrimination value*. The higher the number, the higher and better the discrimination. If, for some reason, more individuals from the lowest-scoring group than from the highest-scoring group select the correct answer, the result would be a *negative discrimination value*. In general, items with a low (or negative) discrimination value should be reviewed and probably edited. A negative discrimination is most likely a miskeyed item. Negative discrimination could also indicate that the item was tricky and that high performers read into the item and missed it, whereas the low performers got it correct. Other common reasons for negative discrimination include multiple correct answers, and distractors that are actually lesser known, special case situations. As with calculating item difficulty, calculating item discrimination can be accomplished through the application of simple mathematical skills.

The item discrimination is calculated (in a simplified manner) by completing the following steps:⁶

1. The instructor identifies the exams with the 10 highest scores and the 10 lowest scores.
2. For each question, the instructor records the number of students in the top group of 10 who answered the question correctly. The instructor does the same for the bottom group of 10 students.
3. The instructor then computes the discrimination index by subtracting the number of students in the bottom group who answered correctly from the number of students in the top group who answered correctly, and dividing by the number of students in each group (in this case 10). For example, if 8 student from the top group answered correctly, and 4 students from the bottom group answered correctly, the discrimination index would be 0.4.

The discrimination index will fall between -1.0 and $+1.0$. The closer the index is to $+1.0$, the more effectively the item distinguishes students who know the material (the top group) from those who do not (the bottom group).

all item types. The stem should be carefully considered to assess for length, ambiguity, and other possible sources for confusion. When evaluating multiple choice items, consideration should also be extended to examine the distractors. Ideally, incorrect responses should be spread across all possible distractors. The instructor should consider item discrimination and the proportion of each distractor chosen to determine next steps. Distractors that are never selected may not be plausible. If knowledgeable test takers are drawn to a particular distractor (shown by a low or negative discrimination), then that distractor may present a possibly correct answer—usually a special case that only advanced students would recognize. Test analysis software can analyze discrimination for each distractor, greatly easing the task of distractor analysis.⁷

It is important for the instructor to note that the item can have an appropriate difficulty level and discrimination index, but if it does not follow the principles of exam item development and construction, it may not be valid and should not be used.

Point Biserial Value

Some test-item analysis programs will report a *point biserial* value for each test item. This statistic is very similar to a discrimination index. However, rather than calculating an index of high overall performance to low overall performance, the point biserial value calculates the statistical correlation between an individual item and the overall score on the exam. Calculation of a correlation coefficient is beyond the scope of this text, but the value is returned by several test analysis programs. Like the item discrimination, the higher the point biserial, the better that item differentiates between those with high overall knowledge and those with low knowledge. Also like the item discrimination, this statistic tends to be low when an item is not very difficult. If an instructor has access to point biserial, it should be used in place of the item discrimination to select those items that require further editing.

Grading Strategies

Appropriate grading strategies are just as important to validity and reliability as the appropriate exam administration and exam content.

Grading Essays

Grading essays and written assignments can be particularly difficult. Because grading essays is inherently subjective, reliability is difficult to ensure. Some suggested strategies that help to improve reliability in the grading of essays are as follows:

- Skim all writing assignments quickly before grading them, to gain an overview of the general level of performance and the range of responses.⁸
- Before the writing assignment is given or the test is administered, the instructor should decide on guidelines for full or partial credit. This is referred to as the *analytic method of grading*. The instructor assigns a number of points to each designated content area. The instructor decides on partial credit for each area and totals the points for an easy grade calculation. It may be useful for the instructor to anchor these points to specific words, phrases, or concepts to help ensure reliability.⁹
- Develop a rubric for scoring essay items, including the characteristics of a correct response and the value of each parameter.¹⁰
- Choose examples of student responses to serve as anchors for different levels of performance. The instructor chooses one student response as an example of a good essay, one as an example of middle performance, and one as a poor example. This approach is referred to as the *global method* of grading. It is generally helpful for instructors to also compare responses with those on an “ideal” paper prepared before the assignment.⁹ Instructors should note that this can be a normative grading scale (discussed in the next section) rather than the more common form of criterion-referenced examination in which performance is compared to an objective standard.
- Grade essay items question-by-question rather than student-by-student. This allows more meaningful comparison of responses between students. Instructors should shuffle the exams between questions to avoid bias in grading caused by student performance on the previous question.⁸ Previous warnings about normative grading also apply to this strategy.
- Avoid judging assignments on the basis of extraneous factors such as illegible handwriting and the use of pen versus pencil. Judge essays on the intellectual quality of the response. Instructors must remember the purpose of the essay question when they are grading.⁶
- If possible, repeat the grading process a couple of days later. Another option is to use multiple graders. Agreement in grades across independent grading sessions supports reliability.⁹

CASE in Point

An instructor includes an essay item relating to the pathophysiology of shock on a module examination. Five points are assigned to the item. She constructs the following rubric to assist her in grading:

- (5 points) Clear description of hypovolemic, distributive, cardiogenic, and obstructive shock with at least two examples of conditions that would cause each
- (4 points) Clear descriptions but missing clearly relevant clinical conditions for some types
- (3 points) Descriptions lack clarity, missing key mechanisms of how perfusion is limited in that type of shock
- (2 points) Descriptions or examples provided for only three of the four types
- (1 point) Descriptions or examples provided for fewer than three of the four types

Norm-Referenced Grading

Normative grading (norm-referenced grading) strategies are those that compare student performance with the performance of other students for assignment of a grade. This is commonly referred to as “grading on the curve.” The result of this strategy is that a set percentage of students receives an “A,” a second group gets a “B,” another group receives a “C,” and some are given a “D” or an “F.” Each student’s grade is determined by the group’s performance—not by comparison with objectives. This method of grading is commonly attacked because it is based on class performance rather than on comparison of performance with objectives. On the other hand, an advantage of normative grading strategies is that the grading strategy automatically compensates for poorly constructed examinations. If a test is very easy, the curve automatically shifts to require a higher passing score. If a test is very difficult, the passing score shifts lower to compensate. This occurs without additional calculation or analysis by the instructor. A number of variations of normative grading strategies include setting a percentage of students that will receive each grade, assigning grade levels based on natural breaks in the distribution, and assigning grade levels based on a normal statistical distribution. Although purely normative strategies are generally considered inappropriate for summative assessment in EMS courses, normative strategies are useful for assigning grades to formative assessments with minimal impact on final grade. Normative approaches, such as grading on the curve, may be useful as an interim

method when using new items in formative assessments such as quizzes.

Criterion-Referenced Grading

Criterion-referenced grading strategies base grade assignment on mastery of course objectives. This approach requires the presence of relatively specific course objectives on which assessments can be based. According to a criterion-referenced strategy, the assessment is drawn from the blueprint, and setting grades is guided by the degree to which objectives are mastered. One example would be that 90% mastery is assigned an “A,” 80% is assigned a “B,” and so forth. This can be based on depth of mastery (90% knowledge of each objective) or breadth of mastery (complete knowledge of 90% of the objectives). A criterion-referenced strategy requires the use of valid and reliable items to ensure fairness in the assessment process. Because the setting of grades does not automatically adjust for difficulty, the instructor must perform additional analysis to set an appropriate passing score.

Setting a Cut Score

The **cut score** is the score required in order to pass a test; it is the passing score. In general, the instructor has two strategies from which to choose when determining the cut score. In the first, the instructor can build the assessments, analyze exam items, and set the passing score based on the difficulty of the exam. In the second, the instructor can first set a passing score, then analyze draft items and construct an examination with difficulty appropriate for the preset passing score. In other words, the instructor can either set the passing score to fit the exam or engineer the exam to fit the passing score. Either option is appropriate. It is inappropriate for students to consider a course with an 80% passing score harder than a course with a 60% passing score, without consideration of the relative difficulty of the examinations.

Many educational institutions set the grade levels and passing scores as part of institutional policy. This fits with a common expectation that 90% = A, 80% = B, 70% = C, 60% = D, and below 60% is failing. Another common expectation is that 70% is passing, with grades interspersed. If instructors are teaching with preset passing scores and grading levels, then they must construct examinations of appropriate difficulty to match this preset passing score. The instructor does this by predicting the difficulty level for each item and computing the average difficulty index for all items on the examination. The instructor can then adjust the examination to match the computed difficulty with the preset passing score.

Setting the Standard: The Angoff Method

An instructor can use a number of methods to predict item difficulty. The most common is the **Angoff method**. This method is commonly used for high-stakes examinations in educational, certification, and licensure settings. The procedure is to first establish a panel of experts. The panel considers the concept of the “minimally competent candidate,” or, in other words, the minimum acceptable level of knowledge. This is not the ideal or average candidate, but the candidate who is barely acceptable. The experts are then asked to estimate the percentage of minimally

competent candidates who would answer that item correctly. This is done first with practice items, where the experts’ estimates can be compared with actual performance of the item. As the experts rate the items, the consensus that is reached by the experts’ estimates forms the Angoff rating. By computing the mean of the Angoff ratings for all items to be included in the exam, the instructor can determine a cut score. Conversion of this predicted cut score into a passing score is a matter of professional judgment for the instructor.

CASE in Point

Setting a Cut Score

An instructor who is teaching a paramedic course at a community college is preparing a module examination for medical emergencies. After collecting and editing a number of examination items, the instructor prepares to predict the difficulty by using the Angoff method. The instructor plans to use the information to set an appropriate passing score.

The instructor contacts four preceptors and three lab assistants who will serve as the expert panel. She starts the process by initiating a discussion on the concept of entry-level competency. She describes the concept in this manner: “The idea is to describe the provider who is barely competent. Not a great paramedic, or even a good paramedic, but instead, the paramedic who has just the amount of knowledge to be considered competent.” She asks panel members to describe in their own words the depth of knowledge required for entry-level competency related to medical emergencies. The discussion continues for a short time until the instructor believes that the panel has reached consensus on the concept.

Next, the instructor distributes a set of examination items that have been used for past courses and for which the actual difficulty level is known. The instructor projects the item, without the answer indicated, and asks the panel, “What percentage of entry-level providers would get this question correct?” After panel members have given their thoughts, she shares with the group the answer to the item. Panel members are then allowed to reconsider their rating. The instructor then shows the group how the item actually performed (the difficulty level of each item from previous

administrations), and the results of each panel member are shown to the group. The panel has a short discussion on the difference between their estimates and the actual performance of the item. This exercise is repeated several times.

After reviewing these practice items, the instructor distributes the ones she will be using for the examination, without an answer key. Each panel member then rates each item as to the percentage of entry-level providers who would answer the item correctly. The answer key is then provided, and panel members are allowed to reconsider their estimate. The instructor collects and averages the results, as shown in **TABLE E**.

The panel of experts has recommended a cut score (minimum passing score) of 70% for this examination. The instructor takes this into consideration as she determines the passing score for the examination. She takes into account that during the practice session, the panel consistently predicted rates of correct responses that were slightly higher than the actual values (in other words, during the practice session, the panel slightly underestimated the difficulty of items). She also considers the potential for error and decides to set the cut score for this examination at 60%.

(Note: In this case, had the instructor been in an institution that mandated by policy a set passing score, the instructor could just as easily use this procedure to predict the item difficulty for each item, then could base item selection on the predicted difficulty to construct an examination of appropriate difficulty for the mandated minimum passing score.)

TABLE E Example of Expert Panel Ratings Used to Determine Cut Score

Item	Panel Member 1	Panel Member 2	Panel Member 3	Panel Member 4	Panel Member 5	Panel Member 6	Panel Member 7	Average
1	70%	80%	80%	80%	70%	80%	75%	76%
2	85%	80%	85%	90%	90%	85%	75%	84%
3	60%	75%	55%	60%	65%	55%	45%	59%
4	75%	70%	70%	80%	85%	65%	80%	75%
5	55%	50%	50%	80%	50%	45%	45%	54%
6	80%	90%	90%	85%	95%	90%	95%	89%
7	50%	50%	60%	70%	55%	50%	40%	54%
8	70%	80%	70%	75%	80%	70%	75%	74%
9	50%	40%	50%	50%	45%	50%	55%	49%
10	80%	80%	80%	80%	90%	85%	85%	83%
(etc.)								
Total (column average)	71%	69%	70%	76%	69%	69%	66%	70%

The Cut Score

The difficulty of the tests an institution uses makes the cut score meaningful. For instance, consider the following two training programs. Program A requires an 80% score to pass the final examination. Program B requires 60% to pass. Program A uses only examination items with an Angoff

rating of at least 90%, with an average Angoff rating of 95%. Program B uses a range of Angoff scores from 40% to 90%, with the average Angoff rating of 60%. Program B is thus a much more challenging program, despite the lower cut score, because it uses much more difficult examinations.

Item-Response Theory

Instead of using a set minimum passing score, the NREMT determines whether a candidate passes the examination by directly assessing the difficulty of the items answered correctly by the candidate. Traditional examinations give all candidates a set number of items of comparable difficulty and compare performance of candidates by the percentage

of items answered correctly. This approach is referred to as a *linear test* and uses classical test theory.

Computer-adaptive testing allows the use of a more precise tool called item-response theory (IRT). Using this approach, a large test bank is established with items of identified difficulty. IRT and the Angoff method are used to

identify item difficulty. The computer adjusts the difficulty of items for the candidate based on the candidate's responses. If a hard item is missed, the next question is slightly easier. If an item is answered correctly, the next is slightly harder. And so on. Each question answered correctly is an indication of the candidate's ability. Once enough items are correctly answered to place the candidate's ability with certainty, the test ends. The more items that are answered, the less the error of measurement. The further the candidate's ability from the competency line, the more measurement error is

allowable to determine with statistical certainty that the candidate is competent (or not competent). Therefore, candidates who are far above or below the competency line will have relatively few questions. Candidates who are near the competency line require many more items to accurately determine whether they meet minimum standards of competence.

For this reason, discussions of scores or test length for computer-adaptive examinations that use IRT are not meaningful.

Summary



Properly constructed written assessments remain the most effective means of easily assessing cognitive objectives. Careful consideration of assessment purpose ensures that the function of the assessment matches the use of written exam items. Different types of written exam tools have varying abilities to assess diverse types of knowledge. Items such as completion and matching are well suited to testing recall. Essay and short-answer items are better for testing analysis and evaluation. Multiple choice and true/false items can test different levels of thinking, but construction of items that evaluate higher-order thinking is challenging. Homework can provide formative assessment. Research projects are another tool that can be used to evaluate higher levels in the cognitive domain. A combination of these different tools provides the instructor with a valid and reliable assessment of a student's mastery of knowledge.

Security of examination materials is important for limited response tools that prove to be valid and

reliable assessments of higher levels within the cognitive domain. Security is not an issue for homework and project assignments, and it is less critical for essay items. Because of the extensive effort needed to properly construct and analyze limited response items, it is necessary that the security of these items be protected. If security is compromised, items that would otherwise test high-end knowledge become items that test only recall. In addition, compromises of examination security may dramatically change the difficulty level and discrimination index of the compromised items.

Limited response items are extremely valuable for the EMS instructor. Testing a large number of cognitive objectives with acceptable reliability and validity requires the use of many more limited response items than essay and short-answer items. This also serves to prepare EMS students for licensure examinations, which use limited response items almost exclusively. Unfortunately, these items become nearly worthless if security is compromised.

Glossary



Angoff method Expert group consensus process to assign item difficulty, in which the group considers the minimum level of acceptable knowledge, then estimates the percentage of minimally competent candidates who would answer the item correctly.

blueprinting Planning the exam to facilitate validity with appropriate level of required thinking, content depth, and breadth.

computer-adaptive testing Type of testing that is able to adjust the difficulty of items for the candidate based on the candidate's responses.

criterion-referenced grading Grading strategies that assign a grade based on mastery of course objectives.

cut score Score required to pass a test; the passing score.

difficulty level Percentage of students who answer each item correctly.

distractor Incorrect answer designed to be a plausible alternative to the correct answer.

item discrimination Degree to which a correct answer for a particular item is associated with high

overall scores on the exam; this is essentially a test of reliability.

item-response theory (IRT) Strategy of measuring a test taker's underlying traits or abilities using performance on different test items, which enables computer-adaptive testing to measure ability much more efficiently than classic tests.

negative discrimination Index that indicates that students who scored well overall did worse on those particular questions than did students who did not score well overall.

normative grading (norm-referenced grading)

Grading strategies that compare student performance with the performance of other students; grading on the curve.

proctoring Act of monitoring the test-taking environment; helps to ensure security of testing materials and to prevent cheating.

stem Part of the item that is first offered, which may be written as a question or as an incomplete statement.

References



- ¹ Hertz, Norman R., and Roberta N. Chinn. 2000. *Licensure Examinations*. Lexington, KY: Council on Licensure, Enforcement, and Regulation.
- ² Margolis, Gregg S., Gabriel A. Romero, Antonio R. Fernandez, and Jonathan R. Studnek. 2009. "Strategies of High-Performing Paramedic Educational Programs." *Prehospital Emergency Care* 13: 505–11. <https://doi.org/10.1080/10903120902993396>.
- ³ Oermann, Marilyn H., and Kathleen B. Gaberson. 2009. *Evaluation and Testing in Nursing Education*, 3rd ed. New York: Springer Publishing.
- ⁴ National Council for State Boards of Nursing. 2018. "NCLEX & Other Exams: What the Exam Looks Like." Accessed December 29, 2018. <https://www.ncsbn.org/9010.htm>.
- ⁵ Cantillon, Peter. 2010. *ABC of Learning and Teaching in Medicine*. Oxford, UK: John Wiley & Sons.
- ⁶ Davis, Barbara G. 2001. *Tools for Teaching*. San Francisco: Jossey-Bass.
- ⁷ Gierl, Mark J., Okan Bulut, Qi Guo, and Xinxin Zhang. 2017. "Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review." *Review of Educational Research* 87, no. 6: 1082–116. <https://doi.org/10.3102/0034654317726529>.
- ⁸ Jacobs, Lucy C., and Clinton L. Chase. 1992. *Developing and Using Tests Effectively: A Guide for Faculty*. San Francisco: Jossey-Bass.
- ⁹ Cashin, William E. 1987. "Improving Essay Tests." *IDEA Paper* no. 17. Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.
- ¹⁰ Johnson, Robert L., James A. Penny, and Belita Gordon. 2008. *Assessing Performance: Designing, Scoring, and Validating Performance Tasks*. New York: Guilford Press.

Additional Resources



- Cantillon, Peter, William Irish, and David Sales. 2004. "Using Computers for Assessment in Medicine." *British Medical Journal* 329, no. 7466: 606–9. <https://doi.org/10.1136/bmj.329.7466.606>.
- Case, Susan M., and David B. Swanson. 2001. *Constructing Written Test Questions for the Basic and Clinical Sciences*, 3rd ed. Philadelphia: National Board of Medical Examiners. Accessed March 1, 2019. https://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf.
- Clegg, Victoria L., and William E. Cashin. 1986. "Improving Multiple-Choice Tests." *IDEA Paper* no. 16. Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.
- Frary, Robert B. 1995. *More Multiple-Choice Item Writing Do's and Don'ts*. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation.
- Haladyna, Thomas M., and Michael C. Rodriguez. 2013. *Developing and Validating Test Items*. New York: Routledge.
- Royal, Kenneth, Marian-Wells Hedgpeth, Jamie Mulkey, and John Fremer. 2016. "The 10 Most Wanted Test Cheaters in Medical Education." *Medical Education* 50, no. 12: 1241–4. <https://doi.org/10.1111/medu.13096>.
- Wendt, Anne, and Lorraine E. Kenny. 2009. "Alternate Item Types: Continuing the Quest for Authentic Testing." *The Journal of Nursing Education* 48, no. 3: 150–6. <https://doi.org/10.3928/01484834-20090301-11>.
- Withers, Graeme. 2005. "Item Writing for Tests and Examinations." UNESCO International Institute for Educational Planning. Accessed March 1, 2019. <https://unesdoc.unesco.org/ark:/48223/pf0000214552>.