Chapter 5 Constructing Tests and Performance Assessments

In a "standards based" approach to education and training, informed by Constructivist theory, assessment informed instruction is the expectation as is continuous improvement. One of the most widely used tools in assessment and evaluation is the traditional or classic classroom achievement test, whether the classroom is on- or offline. These measures are often fraught with reliability and validity problems as the process for constructing such tests is often not followed or misunderstood, thereby introducing significant measurement error into the measurement process. Poor measurement frequently leads to inaccurate data-based inferences, which in turn leads to bad decision-making.

A test is any device (written, observational, or oral) utilized to gather data for assessment and evaluation purposes. The chief assessment device in education and training is the test. Lyman (1998, pp. 21-26) offers one classification taxonomy of various types of tests:

1. Maximum performance tests (MPT): with these, we assume that all examinees will perform their best as all examinees are equally, highly motivated. Examinee performance is influenced by the effects of heredity, schooling or training, and his or her personal environment. A test may fall into more than one classification.
    a. Intelligence tests: These are tests of general aptitude. IQ (Intelligence Quotient) scores will most likely vary according to the intelligence test taken as they do tend to be different. IQ is influenced by previous academic achievement. IQ tests rely on a theory of intelligence, requiring construct validity.
    b. Aptitude (or ability) tests: These tests measure performance potential. Aptitude tests imply prediction and are sometimes substituted for intelligence tests, and used for classification (e.g., ability grouping). Ability test scores are influenced by prior achievement (e.g., reading and math knowledge and skills.). These rely on content validity.
    c. Achievement tests are used to measure examinees' current knowledge and skill level. These tend rely on content validity.
    d. Speeded tests are maximum performance tests where the speed at which the test is completed is a vital element in scoring (e.g., typing test, rifle assembly, foot race, etc.) If a test has a time limit such that virtually all students finish the test, then it is not considered a speeded test and is called a power test. Most achievement tests are also power tests.
    e. Performance tests are designed to require examinees to demonstrate competence by constructing a response. In "traditional" testing, this takes form as a constructed response, i.e., brief or extended essays. A performance assessment may also require a competence demonstration via the construction of a work product; in this instance, detailed product specifications, scoring criteria, and rating or scoring sheets are used.

2. Typical performance test (TPT): These tests include personality, interest, preference, and values tests. They are called instruments, scales, inventories, and indexes. With these types of tests, we assume that examinees will perform typically or "as usual." There is less agreement with what is being measured and what the score means. That is why it is essential that any theory upon which a typical performance test is based be explicitly defined and explained. There is an assumption that an examinee answers test items truthfully. The application of typical performance tests should be cautious in educational applications. These "tests" rely primarily on construct validity.

3. Standardized tests are aptitude or achievement tests which are administered under standard conditions and whose scores are interpreted under standardized rules, typically employing standard scores such as norms, except on criterion-referenced tests. On most standardized tests, the breadth of content coverage is broad and test items are written to maximize score variability. Thus, items which are too hard or too easy are not used in constructing the test. Standardized achievement and aptitude tests rely on content validity.

4. Informal tests are those typically intended for a localized purpose, such as the simple classroom achievement test.

We focus on achievement testing which is used most to measure student or trainee mastery or command of a subject area (e.g., math, biology, safety laws, customer service policies, etc.). The achievement test construction process is seven (7) steps:

1. First, when planning the assessment or test, consider examinees' age, stage of development, ability level, culture, etc. These factors will influence construction of learning targets or outcomes, the types of item formats selected, how items are actually written, and test length.

2. Second, it is necessary that the content, intellectual or thinking skills (see Appendix 5.1) psychomotor skills, and/or attitudes to be assessed are fully and clearly identified and written. This is usually done through learning outcomes or standards and benchmarks. Some write lesson plans at this step.

3. Third, the test item specifications are written. Item specifications intended for a state-, provincial- or nation-wide tests will be more detailed than will a common departmental or classroom examination, which will probably be just the learning outcome/target benchmarks and possible test item formats. Detailed item specs are written if different item writers construct tests for different examinee pools, e.g., school districts permitting each school to write its own unit examinations.

4. Fourth, a table of specifications (or test blueprint) is developed which integrates the content, psychomotor skills, and/or attitudes to be assessed with the intellectual skills and selected test item formats, including the number of items per learning target benchmark. See Table 5.1.

5. Fifth, once steps 1 to 4 are completed, the plan is initially reviewed by knowledgeable colleagues to assess accuracy, clarity, continuity, and "fit." The reviewers also try to detect bias. This is an iterative process which should be engaged until there is general agreement that bias isn't present. This step is critical to establishing content validity.

6. Sixth, once a consensus has been reached, items are written, based on item writing guidelines, and revised until there is agreement that each item (a) meets its item specification, (b) conforms to item writing guidelines, (c) "fits" the test blueprint, and (d) no bias is present; the determination is done by the same or similar subject matter experts completing Step 4, above. Step 6 is again critical to establishing content validity. The outcome of Step 6 is an initial version of the test.

7. Seventh, once the initial test is written, it is typically pilot tested with examinees similar to the intended target audience and evaluated statistically (using item analysis indices; see Part IV of the chapter). The statistical evaluation provides guidance for the revision or rejection of poorly functioning items. Pilot more test items than will be actually needed. The outcome of Step 7 is the final version of the test.

In this chapter, we will first examine "mechanics of achievement test construction" (e.g., test planning, building learning targets, test blueprinting, test item type and item format selection), writing select response items (multiple-choice, true/false, matching, completion and short-answer) and supply response items (brief and extended response) is discussed. Next, statistical strategies for improving test items are presented, followed by performance assessment. Ethical strategies for preparing examinees for a testing session are presented in Appendix 5.6.

I. The Mechanics of Achievement Test Construction

A. Learning Standards, Targets, or Outcomes: The Assessment Drivers

1. In education and training, learning outcomes, also called standards or targets, are framed to specify the content, skill, or attitude to be measured. These standards drive curriculum writing, teaching and learning, test construction, and testing (assessment).
    a. Learning outcomes or standards have also been referred to as content standards, performance standards, or behavioral objectives. Here, these terms are used interchangeably.
    b. There are three (3) types of learning standards, targets, or outcomes.
        (1) Types of Learning Standards
            (a) Attitude standards state explicitly what attitudes, based on defined values, the faculty expect students to hold and articulate as a function of program or course enrollment.
            (b) Content standards express explicitly what content students are expected to know.

(c) Skill standards state very clearly the specific skills students are expected to have mastered at a specified performance level.

(2) It is often easy to confuse content and skill standards.

(a) More specifically, content standards specify declarative knowledge, e.g., mathematical rules, statistical formulas, important historical facts, grammar rules, or steps in conducting a biology experiment, etc.

(b) Skill standards specify procedural or conditional knowledge, e.g., conducting statistical or mathematical operations based on formulas or mathematical rules; interpreting or explaining historical facts or analyzing historical data; correcting a passage of text for poor grammar; or conducting a biology experiment.

(c) The key difference between content and skill standards is that with content standards, students are required to possess specific knowledge; skill standards require students to apply that knowledge in some expected fashion at an expected level of performance.

c. In crafting learning standards, targets, or outcomes, Oosterhof (1994, pp. 43-47) suggests the writer consider:

(1) Capability. Identify the intellectual capability being assessed. Performance taxonomies such as Bloom, et al. or Gagne are helpful here.

(2) Behavior. Indicate the specific behavior which is to be evidence that the targeted learning has occurred. The behavior should be directly observable, requiring no inference.

(3) Situation. Often, it is helpful to specify the conditions under which the behavior is to be demonstrated. Describe the circumstances the behavior is to be demonstrated.

(4) Special Conditions. Depending on the circumstances, one may need to place conditions on the behavior, e.g., name a letter or word correctly 80% of the time in order to conclude that the targeted learning has occurred.

2. There are several models for framing these intended outcomes or standards; we briefly examine two and then integrate these two approaches into one, guided by Oosterhof's (1994, pp. 43-47) suggestions.

a. Mitchell (1996) offers the following Taxonomy and Definitions

(1) Content (or academic) standards. These standards identify what knowledge and skills are expected of learners at specified phases in their educational progression.

(2) Performance standards. Performance standards have levels, e.g., 4, 3, 2, 1; exceeds expectations, meets expectations, or does not meet expectations; unsatisfactory, progressing, proficient, or exemplary, which are intended to show degree of content mastery.

(3) Opportunity to learn standards. There are instances where enabling learning and/or performance conditions are identified to ensure that

learners have a fair chance to meet Content and Performance Standards.

b. The State of Florida (1996, pp. 28-30) developed the following taxonomy:
   (1) A Strand is a label (word or phrase) for a category of knowledge; consider it to be a general content (knowledge and/or skill) organizer.
   (2) A standard is a general statement of expected learner achievement.
   (3) A Benchmark describes in, in a series of statements more precisely what a learner is expected to know and/or be able to do at the end of a training or educational unit or program. When all associated benchmarks are accomplished, the standard is met.

3. The Integrated Model
   a. A standard is typically composed of five elements.
      (1) The first element states "who is to do something", usually a student.
      (2) The second element is an action oriented verb (e.g., articulate, describe, identify, explain, analyze, etc.); it is at the verb level that intellectual skill taxonomies (e.g., Bloom) exert their influence. For example, Quellmalz outlined five cognitive functioning levels (Bloom's equivalencies are noted), which are:

Quellmalz (1987)	Bloom, et al. (1956)
Recall	Knowledge & Comprehension
Analysis	Analysis
Comparison	Synthesis
Inference*	Application & Synthesis
Evaluation	Synthesis & Evaluation

*(deductive & inductive reasoning)

For instance, according to Stiggins, Griswold, and Wikelund (1989), verbs associated with inference include generalize, hypothesize, or predict. Within Quellmalz's "comparison level," one could use the words compare and contrast. We can examine intellectual skills taxonomy to find the most precise verb to describe what we want an examinee to know, do, or feel. There are other intellectual skill taxonomies such as Canelos (2000) and Gange's (1985); others are presented in Appendix 5.1. The key point is to identify a useful one and use it to guide learning standard, target, or objective writing.

      (3) The third element describes under what condition(s) the student is to demonstrate something (e.g., fully, briefly, clearly, concisely, correctly, accurately, etc.).
      (4) The fourth element specifies what the student is to demonstrate (e.g., algebra calculation, leadership theories, decision-making models, etc.)
      (5) The fifth element (optional) describes the medium in which the demonstration is to take place, e.g., in written or oral form, via examination, case report, or portfolio, etc. We don't recommend using

the fifth element as assessment options may become too limited. Two sample standards are:
(a) The student will accurately compute algebraic equations.
(b) The student will accurately and concisely describe modern leadership theories.

b. Construct an Operational Definition, called a benchmark for each Attitude, Content, and/or Skill Standard
   (1) Standard operational definitions are constructed through benchmark statements. A benchmark, in plain English, is a specific action oriented statement which requires the examinee to do something. When a student has achieved each relevant benchmarks, the standard is met. Benchmarks further define the relevant attitude, content, and skill domains. Illustrative benchmarks are:
      (a) The student will accurately compute algebraic equations.
         [1] The student will correctly compute linear equations.
         [2] The student will correctly compute quadratic equations.
         [3] The student will correctly compute logarithms.
      (b) The student will accurately and concisely describe modern leadership theories.
         [1] Describe content and process motivation theories citing education leadership examples.
         [2] Describe leadership trait theories of leadership, respecting assumptions, elements, research, and education application.
         [3] Describe leadership styles and situational models in terms of assumptions, characteristics, and education application.
   (2) Benchmarks are very useful in framing examinations, assignments, etc. Test items are written based on specific benchmarks. As with standards, subject matter experts (SME's) should agree on the appropriateness of each benchmark and its association with its specific standard. Further, it is not usually necessary to stipulate a performance level as in percent correct; the school's grading scale or the organization's performance assessment system should suffice.

c. Once a sufficient number of standards, in depth and breadth, have been drafted, units, courses, degree programs, etc. are then constructed around relevant "bundles" of standards. Such "bundles" of relevant attitude, content, and skill standards define what is to be taught (i.e. the curriculum-- also called a domain), learned, and tested.
   (1) Attitude standards can be assessed by surveys; see Chapter 4.
   (2) Content standard mastery can be studied through examinations which require the student to demonstrate his or her knowledge.
   (3) Skills can be tested by examinations using application oriented item formats (e.g., brief and expended responses or solving problems), projects, portfolios, or cases, etc.

d. The five part model presented here can also be used to evaluate standards written by others. Regardless of the approach employed, each standard should meet at least the first four components, be developmentally appropriate, and ensure that students have had the opportunity to learn, develop the desired attitude, and/or acquire the specified skill(s).

B. Constructing a Test Blueprint

1. A table of specifications sorts the performance standards, i.e., content and/or skills, by intellectual skills to be performed. The number of test items is inserted in the appropriate table cell. An example is presented Table 5.1, where the number of test items per learning benchmark and intellectual skill is presented. An alternative is to list the actual test item number instead of just the number of test items.

2. The more important the content, the greater the number of test items. To arrive at these numbers, the test developer will typically
   a. Determine the total number of items to be included. This number is influenced by the testing time available, the test environment, developmental status of the examinees, and size of the content and/or skill domain to be tested.
   b. Test items are allocated to each learning target, objective, or outcome; more critical standards are allotted more items.

3. Next, test items are sorted across each learning target or benchmark, by intellectual skill. Test item classification guidelines are:
   a. Determine exactly what the action verb in the Learning Target benchmark asks or requires of the examinee.
   b. Ensure the test item asks or requires the examinee to know or do what is specified in the benchmark. The test item and benchmark must match.
   c. Examine the benchmark's action verb, to classify the item.
      (1) If the examinee is to just recall or know data or information, classify under "Knowledge," which typically is declarative knowledge.
      (2) If the examinee is to recall or identify data or information in a different manner than he or she first learned, classify under "Comprehension," which typically is declarative knowledge.
      (3) If the examinee is to apply (i.e., use) the knowledge or skill learned, then classify under "Application," which typically requires procedural and/or conditional knowledge.
      (4) If the examinee is to analyze (i.e., take apart to thoroughly examine) an idea, recommendation, poem, report, argument, or an interpretation of a dataset or what someone else produced, classify under "Analysis."
      (5) If the examinee is to write a poem, article, term paper, play or perform an artistic routine, or paint a painting, or prepare a photography exhibit, classify under "Synthesis." Of all of Bloom's levels, creativity is most associated with "Synthesis," followed by "Analysis." An example is the construction of a term paper where "Analysis" is

demonstrated by reading articles on a topic and then assembling the learning from the articles into a coherent paper (Synthesis).

d. Carefully select the best action verb which requires the examinee to know the content, perform the skill or display the attitude intended. There are many sources from which to select the best action verb, e.g., the Internet, colleagues, or the synonym list found in most word processing programs.
e. If there is a question about how to classify an item, consult an experienced colleague. Sometimes there is item classification disagreement, due to differing perspectives; try to reach consensus. The test designer usually makes the final classification decision.

4. Application: Table 5.1
   a. Presented in Table 5.1 is a test blue print for an end-of-term test in grades 3-5 language arts. The performance standards are benchmarks drawn from Standard 2 of the 1996 Florida Sunshine State standards for Language Arts. The benchmarks are presented in the far-left column and the intellectual skills are presented in columns to the right. The numbers in table cells are the number of items per intellectual skills.
   b. Upon reviewing the table, the reader sees that this instructor selected (LA.A.1.2.2) as more critical than the others.
   c. For (LA.A.1.2.2), there are 8 knowledge items, which require simple recall of information. Knowledge of phonetic sounds and word structure is critical if a student is to be able to apply this knowledge to construct meaning from various texts. The types of knowledge to be demonstrated drive which item formats are selected. Possible item formats are simple multiple choice, matching, fill-in-the-blank, or true/false.
   d. Next, the examinee should be able to demonstrate a comprehension of what is read in the form of retelling and self-questioning. Comprehension involves translation, interpretation, and/or extrapolation. In the present instance, we would focus on interpretation and extrapolation. Possible item formats are multiple choice, short response, or extended response items.
   e. The test developer elected to test at the application level. To test application, the test developer can craft one table and one graph. Several multiple choice questions to assess an examinee's ability to comprehend and apply data from the table and graph can be constructed. An alternative is to have the student write several sentences as to the meaning and potential use of the data presented by table and graph.
   f. Students will also need to selected texts or illustrations; he or she could be asked to construct a statement explaining the texts and predicting a potential outcome from that meaning. Answers could be selected from multiple choice items and/or supplied as in a written paragraph or two. Since there is only two hours for the examination and more time is required to demonstrate higher order intellectual skills (i.e., analysis, synthesis, and evaluation), there are fewer items included on the test.

Table 5.1
Test Blue Print for End of Term Test on Language Arts Strand A: Reading
Standard: 2. The student writes uses the reading process effectively, Grades 3-5

Benchmark	Knowledge	Comprehension	Application	Analysis	Synthesis	Total
The student will…	5[a,b]	6[d,e]	1[f]			12
Uses a table of contents, index, headings, captions, illustrations, and major words to anticipate or predict content and purpose of a reading selection. (LA.A.1.2.1)						
Selects from a variety of simple strategies, including the use of phonetics, word structure, context clues, self-questioning, confirming simple predictions, retelling, and using visual cues to identify words and construct meaning from various texts, illustrations, graphics, and charts. (LA.A.1.2.2)	8[a,d]	4[c]	2[f]	1[g]		15
Uses simple strategies to determine meaning and increase vocabulary for reading, including the use of prefixes, suffixes, root words, multiple meanings, antonyms, synonyms, and word relationships. (LA.A.1.2.3)	6[a,c]	3[a]	3[e]			12
Clarifies understanding by rereading, self-correction, summarizing, checking other sources, and class or group discussions. (LA.A.1.2.4)			5[f]			5
Item Totals	19	13	11	1		44

Note: The source of the example is: *Florida Curriculum Framework: Language Arts PreK-12 Sunshine State Standards and Instructional Practice* (Florida Department of Education, 1996, pp. 36-37). It is increasingly common to insert the actual test item numbers once the test has been finalized as a quality control strategy. Item Formats: [a]multiple choice, [b]true/false, [c]matching, [d]fill-in-the blank, [e]completion, [f]brief response, [g]extended response.

D. Directions and Arranging Test Items

1. Directions
   a. Directions should be clear, concise, and to the point.
   b. Directions should include: how to record answers; the time available; the points associated with specific subtests or items; what to do when the test is completed; and explain the permissible use of "scratch" paper to show computations, if allowed.
   c. Keep the directions and associated items on the same page, even if the directions need to be repeated on subsequent page(s).

2. Arranging Items on the Test Instrument
   a. Include an "Ice breaker" item which virtually all examinees will answer correctly to build confidence. Since the purpose of test items is to identify those who can answer the item correctly, not all examinees should be able to answer every item on the test.
   b. Group similar content together on the testing instrument. Use items designed to measure important content, as testing time is almost always limited.
   c. Use items of appropriate known difficulty levels (see Part IV of this chapter) when possible.
   d. Don't break items across a page.
   e. Keep charts or figures, pertaining to an item, on the same or next page.
   f. If using a computer printer, use a consistent font or font size.
   g. If students are required to supply an answer, provide enough room for the answer.

E. Timing and Testing

1. Testing time limits are useful because:
   a. Examinees complete the test more rapidly.
   b. Examinees learn to pace themselves.
   c. Examinees may be more motivated.
   d. A time limit can be selected so that most examinees will complete the test.

2. For speeded tests, an examinee's score is materially influenced by the number of items answered correctly within a specified time limit. Thus, item order and format are important. Care must be taken to ensure that item order and format are consistent with the purpose and cognitive (e.g., intellectual or thinning) skill level(s) to be measured. Before time limits are set, there should be pilot-testing to ensure that the time limit is sufficient to accomplish the purpose of the test. Speeded testing is used in emergency skills testing, crisis management training and other circumstances where responding quickly and correctly in the shortest or least amount of time is essential.

3. For power tests (i.e. achievement testing), an examinee's score is influenced by item format and item difficulty as well as the cognitive skill(s) to be assessed. If used, the set time limit should be sufficient so that virtually all examinees will complete the test. Recommended time limits are:
   a. Simple true/false or matching: 15-20 seconds for each item.
   b. Complex true/false or matching and one or two word completion: 20-30 seconds for each item.
   c. Simple multiple choice: 25-30 seconds for each item.
   d. Complex multiple choice: 45-60 seconds for each item.
   e. Brief or restricted response essay: 10 minutes for each item.
   f. Remember, the time required for examinees to answer test items is influenced by reading skill, item format, item length, and cognitive skill(s) being tested. Items which require computations will take longer. The above time limits are intended only to serve as a guide.

F. Interpreting Test Scores

1. Test items are proxy measures of an unobservable psychological construct or trait; declarative, procedural or conditional knowledge; and/or psychomotor skill. Test items also require examinees to use intellectual and/or thinking skills (see Appendix 5.1). Measuring psychological attributes, such as ability or achievement is often not directly possible; that is why test items are written, so that inferences can be made from the examinee's behavior (answering the test item or performing the skill as requested).
   a. Test items consist of a stimulus which is intended to prompt a prescribed or expected answer. A correct answer or endorsement to a test item is seen as an indicator that the examinee has the attribute, knows the knowledge, or can perform skill taught.
   b. Test item formats include: multiple choice, true/false, completion, fill-in-the blank, short answer, check-list, and essay. Each item type has its strengths and weaknesses and should be selected carefully. Guidelines for writing each type of test item are provided in Part III of this chapter.

2. There are three types of test items; the key difference is the test item's purpose.
   a. Mastery items measure essential minimums that all examinees should know; measure memorization of facts or simple computations (e.g., Bloom, et al.'s knowledge or comprehension skill levels); and are commonly used in licensing or certification tests.
   b. Power items are designed to measure what typical examinees are expected to know; these items may range in difficulty form very easy to very hard, depending on the test purpose and the content, skills, or attitudes being measured. Power items are commonly found in achievement tests.
   c. Speed items are designed to assess higher level concepts and skills (e.g., Bloom, et al.'s application, analysis, synthesis, or evaluation) expected of the most able examinees. Speed items are found on tests of mental abilities (e.g., IQ tests). Don't confuse these item types with speeded tests where

the strictly observed time limit is part of the measurement process (e.g., decision-making accuracy in an emergency situation by a paramedic, firefighter, or police officer).

3. Test items must have two (2) very important characteristics: test items are unidimensional and logically independent
   a. Items are Unidimensional
      (1) Each test Item must measure one attribute, e.g., knowledge, attitude, or psychomotor skill, etc.
      (2) If it were possible, in theory, to write every possible test item for an attribute (e.g., knowledge, skill, or attitude), they would fully describe every single aspect of that attribute. There would be too many test items; so, we write test items to measure the most important attribute(s) of the content, skill or attitude of interest.
      (3) This unidimensional assumption is critical for two reasons: (a) it permits a more accurate interpretation of a test item and (b) allows the explanation that a single trait (e.g., knowledge, skill, or attitude) accounts for an examinee responding correctly to the test item.
   b. Items are logically independent
      (1) An examinee's response to any test item is independent of his or her response to any other test item. Responses to all test items are unrelated, i.e., statistically independent.
      (2) Practical implications for writing test items are: (a) write items so that one item doesn't give clues to the correct answer to another item and (b) if several items are related, e.g., to a graphic (picture, table or graph), they should be written so as not to "betray" one another but to test different aspects of the graphic. See Interpretative Exercise in the discussion on writing multiple choice items and Appendix 5.2.

4. Test items are either dichotomously or polychotomously scored.
   a. Dichotomously scored test items are scored as correct or incorrect, right or wrong, etc. This scoring applies to multiple-choice; true/false; matching; completion; and fill-in-the-blank item formats. These items are usually worth at most a few points each.
   b. In polychotomously scored items, there is more than one correct response for a test item or partial credit, i.e., differing point awards are given (e.g., 7 of 10 possible points). This scoring is typically applied to brief/restricted or extended response items, which are usually weighted with more points than select response items. A scoring rubric is used to guide point awards.

5. Allocating Test Points
   a. For a traditional classroom achievement test, usually given at the end of a unit or module, once the test blue print and item formats are selected, points are allocated. Points are necessary to compute, report, and interpret test scores, as well as any subsequent statistical analysis.
   b. One must know how scores are to be interpreted to build a scoring plan that will permit the desired interpretation which is usually either pass/fail or performance segmented as in a grading scale (e.g., "A," "B," "C," etc.). In short, the examiner must know how he or she wants to interpret the data (test score) before the test is constructed.
   c. The most critical knowledge or skills are weighted with the most points, followed by less critical knowledge and skills, receiving fewer points.
      (1) Let's look at an example drawn from the test blueprint presented in Table 5.1. In this example, we will consider each standard as a separate subtest.
      (2) Assigning Points to Items (Table 5.2)
         (a) Brief (or restricted) and extended response items were weighted at 5 and 8 points respectively due to the higher difficulty level and the higher order intellectual skills required to correctly answer the item. Since partial credit is awarded for these items, a scoring rubric to guide awarding points is required.
         (b) Multiple choice and completion items were weighted at 3 points each as more is the required of the examinee to answer the item correctly than let's say a true/false item.
         (c) Fill-in-the blank and matching items were scored at 2 points each as the context of the sentence or clue may trigger examinee memory, thus making it easier for the examinee to recall the correct answer.
         (d) True/false items were weighted at one point each as the examinee is expected to simply recall the correct information.
      (3) There are 4 subtests for a total of 129 points. The examiners weighted Subtest 3 (Standard L.A.A.1.2.2) at 46 points as they thought it assessed more important skills than the other three.
      (4) For interpretative purposes, we can divide earned points by possible points to arrive at a percentage level of mastery (for tests not totaling exactly 100 points) on each subtest and the entire unit for the examine group. Let's suppose the class scored as shown in Table 5.3, using the class average scores on each subtest and the total test score.
         (a) If 80% was the passing or "cut" score, the class as a whole passed the unit examination. However, group/class remediation appears necessary on the Subtest A (Learning Standard L.A.A.1.2.1) content, since the pass rate was 72% (18/25).
         (b) Individual examinee performance can be reviewed and interpreted in the same manner. First, the total test score is examined. Second, each subtest score is examined; if the total test score and individual subtest scores are acceptable, then no further investigation is

warranted. However, if ether the total test or any subtest score is below the passing or "cut" score, then individual items need to be reviewed to determine what re-teaching or remediation is needed for each particular examinee.

Table 5.2
Language Arts Unit 3 Test Scoring Plan

Standard	Item Format	Item Points	Subtest Points
L.A.A.1.2.1	Multiple Choice	3	x 2 items = 6
(Subtest 1)	True/False	1	x 3 items = 3
	Fill-in-Blank	2	x 3 items = 2
	Completion	3	x 3 items = 9
	Brief Response	5	x 1 item = 5
			Total = 25 points
L.A.A.1.2.2	Multiple Choice	3	x 4 items = 12
(Subtest 2)	Matching	2	x 4 items = 8
	Fill-in-Blank	2	x 4 items = 8
	Brief Response	5	x 2 items = 10
	Extended Response	8	x 1 item = 8
			Total = 46 points
L.A.A.1.2.3	Multiple Choice	3	x 6 items = 18
(Subtest 3)	Matching	2	x 3 items = 6
	Completion	3	x 3 items = 9
			Total = 33 points
L.A.A.1.2.4	Brief Response	5	x 5 items = 25
(Subtest 4)			**Total =25 points**

Table 5.3
Language Arts Unit 3 Subtest and Total Test Scores

Subtest	$\bar{x}$ Points Earned	Points Possible	Percent Mastery
A (L.A.A.1.2.1)	18	25	72%
B (L.A.A.1.2.2)	39	46	85%
C (L.A.A.1.2.3)	28	33	85%
B (L.A.A.1.2.4)	20	25	80%
Total Test Score	105	129	81%

d. For a direct performance assessment, each assignment has a task description and scoring rubric so students clearly understand what is expected of them.
   (1) An accurate assignment score or grade assumes a clearly understood task description and consistent application of the scoring rubric for each assignment.
   (2) An assignment score or grade should not be based on anything other than the examinee's performance using the task description and scoring rubric. Don't include participation, effort, or attendance.

(3) Points are allocated in the same manner as a traditional classroom test with the most critical information receiving the highest point allocation and less import receiving fewer points. See Appendices 5.2, 5.4 and 5.5, which contain task descriptions and scoring rubrics.

6. Assigning a Course or Training Program Grade
   a. A course, consisting of several units or modules with multiple assignments and/or tests, may present points possible in a table, such as Table 5.4.
      (1) The 2 performance tasks are weighted the most points as they require the demonstration of higher order intellectual skills, skillful application of the content, and procedural knowledge.
      (2) Discussion prompt answers are a maximum of 500 words each requiring references; they must evidence at least one higher order intellectual skill. Answers are graded with an analytical rubric, which is discussed in Part V of this chapter.
      (3) Students are required to comment on other student discussion prompt answers or comments. Comments are scored holistically.
      (4) The mid-term and final exam assesses knowledge, comprehension, and limited application of critical course knowledge and skills.
   b. Points possible are then usually "segmented," based on point levels or percentages, into performance bands, with qualitative labels; see Table 5.5. Be clear and consistent on the "rounding" rules, 85.8% becomes 86%.

Table 5.4
Course Assignment & Test Point Weights

Assignment	**Points**
Traditional Classroom Test Construction Task	224
Direct Performance Assessment Construction Task	160
Discussion Prompt Answers (8 x 12 points x 2)	192
Discussion Answer Comments (16 x 4 points x 1.5)	96
Mid-term and Final (2 @ 75 points each)	150
Total Points	822

   c. Combing the information from Tables 5.4 and 5.5, a student earning 739 points is assumed to have "mastered" 90% (actually 0.899%) of the content (specified by the learning outcomes or targets) for a performance indicator of an "A-" or an "Excellent" performance.

Table 5.5.
Course Grading Scale

Grade	**Percentage**	**Meaning**
A	95-100%	Exceptional
A-	90-94%	Excellent
B+	87-89%	Very Good
B	83-86%	Good
B-	80-82%	Fair
C	75-79%	Marginal
F	≤75%	Failure

Note. Saint Leo University (2010). *Graduate academic catalog 2010-2011*. St. Leo, FL: Author.

II. Constructing Select Response Items to Assess Knowledge, Comprehension, & Application

A. General Item Writing Guidelines

1. Make the difficulty and complexity of the item appropriate to the examinee's level. Ensure that the item reading level is appropriate for examinees.
2. Define the item as clearly as possible.
3. Write simple, straightforward items. Use the most precise words to communicate the intent of the item.
4. Know the students' mental processes and frame the item accordingly.
5. Vary item complexity and difficulty.
6. Make items as independent from each other, except when used in a series.
7. Avoid negatively phrased items; avoid double negatives.
8. Use correct grammar, syntax, spelling, etc.
9. When writing specific item formats, review the unique item construction guideline for each format to ensure conformity.
10. Ensure that items do not contain language which is racially, ethnically, religiously, or economically biased.
11. Select response items usually measure Bloom's "Knowledge" and "Comprehension" levels. Supply response items measure the higher intellectual skills, "Application," "Analysis," "Synthesis," and "Evaluation."
12. Avoid clues that Airasian (1997, p. 177) calls specific determiners. These are words which tend to "give away" the correct answer. For example, words such as "always", "all", or "none" tend to be associated with false "true/false" items.

B. Writing Multiple-Choice Items

1. Each multiple-choice item has a stem which presents the situation, problem, definition, etc. Then there are answer choices (also called options or foils) which include the correct answer and plausible wrong answers, called distracters or foils. There are several multiple choice item variations:
   a. Correct Answer: Only one answer is correct.
      (1) Correct Answer

      ___In times of war in the early Roman Republic, the two consuls stepped aside in favor of one person who would be responsible for making decisions quickly. That person was a —

      a. General c. Dictator
      b. Czar d. Tyrant

      (2) Correct Answer

      ___What is 68 x 22 = _____

      a. 149 c. 1,496
      b. 1,469 d. 4,196

      (3) Answers: "c" & "c"

b. Best Answer: Examinees select the "best" response option that "fits" the situation presented in the item stem.

(1) Best Answer

___This culture developed an accurate calendar. They built steep temple pyramids and used advanced agricultural techniques. They developed a system of mathematics that included the concept of zero. They were located mostly in the Yucatán Peninsula. They ruled large cities based in southern and southeastern Mexico, as well as in the Central American highlands. This passage best describes the ancient.

a. Aztecs c. Mayas
b. Incas d. Olmecs

(2) Answer: "c"

c. Rearrangement: This item is can be used to assess examinee procedural knowledge or comprehension.

(1) Rearrangement

There are several preconditions which must be satisfied before a test is constructed. Read each statement listed below and place each in the correct order of operation, by using the codes presented below.

a. First Step c. Third Step
b. Second Step d. Fourth Step

____1. The intellectual skills, attitudes, and/or psychomotor are identified.
____2. The test item formats are considered and selected.
____3. A table of specifications (or test blueprint) is developed.
____4. Items are crafted.

(2) Answer: 1. a; 2. b; 3. c; 4. d

d. Substitution Item: In this variation, the item stem contains a blank or blanks. The examinee then selects from the response options, either the correct or best response for the stem; the purpose is usually intended to make the stem a correct or true statement. The item is widely used to assess examinee comprehension.

(1) **Directions**. Please the letter which represents the word which correctly completes the statement in the corresponding blank.

Reliability and Validity Relationship

a. Concurrent c. Content e. Internal Consistency
b. Construct d. Test-retest f. Alternate Forms

At a minimum all tests must have (a) _____validity and (b) ____ reliability.

(2) Answers: 1c; 2e

e. Analogy

(1) A multiple choice item can be written in the form of an analogy.

(2) ___Hieroglyphics is to Egypt as cuneiform is to —

a. Phoenicia c. Persia
b. Sumer d. Crete

(3) Answer: b

2. Strengths and limitations of multiple choice items are:
   a. Strengths
      (1) Simple and/or complex learning outcomes are measured.
      (2) Highly structured and clear tasks are provided.
      (3) These items are highly efficient in assessing substantial content domains quickly.
      (4) All responses, even incorrect endorsements, provide useful item revision and examinee performance information.
      (5) Items are less susceptible to guessing than matching or true/false items.
      (6) Scoring is simple, quick, and objective.
   b. Limitations
      (1) Constructing high quality multiple choice items is labor and time intensive.
      (2) Writing plausible foils is often difficult.
      (3) Writing items to measure higher order intellectual skills is very difficult and when successfully done provide only a proxy measure of such skills.
      (4) Like most test items, the multiple choice format can be influenced by reading skills.

3. A review of the test item design literature (Oosterhof, 1994, pp. 33-147; Gronlund, 1998, pp. 60-74; Popham, 2000, pp. 242-250) and the authors' reveal these item writing guidelines.
   a. Learning outcomes should drive all item construction. There should be at least two items per key learning outcome.
   b. In the item stem, present a clear stimulus or problem, using language that an intended examinee would understand. Each item should address only one central issue.
   c. Generally, write the item stem in positive language, but ensure to the extent possible, that the bulk of the item wording is in the stem.
   d. Underline or bold negative words whenever used in an item stem.
   e. Ensure that the intended correct answer is correct and that distractors or foils are plausible to examinees who have not mastered the content.
   f. Ensure that the items are grammatically correct and that answer options are grammatically parallel with the stem and each other.
   g. With respect to correct answers, vary length and position in the answer option array. Ensure that there is only one correct answer per item.
   h. Don't use "all of the above," and use "none of the above" unless there is no other option.
   i. Vary item difficulty by writing stems of varying levels of complexity or adjusting the attractiveness of distractors.
   j. Ensure that each item stands on its own, unless part of a scenario where several items are grouped. Don't use this item format to measure opinions.
   k. Ensure that the arrangements of response options are not illogical or confusing and that they are not interdependent and overlapping.

l. Avoid clues that enable examinees to eliminate incorrect alternatives or select the correct answer without really knowing the content. Common clues include:
   (1) Restating the text or lecture note answer verbatim or nearly verbatim.
   (2) Using implausible distractors.
   (3) Writing the correct answer so that it is longer than other plausible distractors.
   (4) Writing distractors which are so inclusive that an examinee is able to exclude other distractors or distractors which have the same meaning.
   (5) Absolutes such as "never", "always", etc. are associated with false statements. Their use is a clue to examinees that the alternative is most likely incorrect.

4. Specific Applications
   a. Best Answer Items: In this scenario, the examinee is required to select the best response from the options presented. These items are intended to test evaluation skills (i.e., make relative judgments given the scenario presented in the stem).
   b. To assess complex processes, use pictorials that the examinee must explain.
   c. Use analogies to measure relationships (analysis or synthesis).
   d. Present a scenario and ask examinees to identify assumptions (analysis).
   e. Present an evaluative situation and ask examinees to analyze applied criteria (evaluation).
   f. Construct a scenario where examinees select examples of principles or concepts. These items may either measure analysis or synthesis, depending on the scenario.

5. The Interpretive Exercise (IE)
   a. A preferred item format for assessing application, analysis, synthesis, and evaluation is the Interpretive Exercise. See Appendix 5.2**.**
      (1) In IE, based on "introductory information", several related items are used to assess mastery of a particular intellectual skill expression.
      (2) This assessment strategy is used to assess other complex skills such as reading ability, comprehension, mathematical thinking, problem solving, writing skills, graph and table interpretation, etc.
   b. IE can be used to assess whether or not examinees can
      (1) Recognize and state inferences;
      (2) Distinguish between warranted/unwarranted generalizations;
      (3) Identify and formulate tenable hypotheses;
      (4) Identify and interpret relationships;
      (5) Recognize assumptions underlying conclusions;
      (6) Formulate valid conclusions and recognize invalid ones;
      (7) Recognize relevance of information; and
      (8) Apply and/or order principles.

c. Advantages
   (1) Interpretive skills are critical in life.
   (2) IE can measure more complex learning than any single item.
   (3) As a related series of items, IE can tap greater skills depth and breadth.
   (4) The introductory material, display or scenario can provide necessary background information.
   (5) IE measures specific mental processes and can be scored objectively.

d. Limitations
   (1) IE is labor and time intensive as well as difficult to construct.
   (2) The introductory material which forms the basis for the exercise is difficult to locate and when it is, reworking for clarity, precision, and brevity is often required.
   (3) Solid reading skills are required. Examinees that lack sufficient reading skills may perform poorly due to limited reading skills.
   (4) IE is a widely used proxy measure of higher order intellectual skills. For example, IE can be used to assess the elements of problem solving skills, but the extent to which the discrete skills are integrated.

e. Construction Guidelines
   (1) Begin with written, verbal, tabular, or graphic (e.g., charts, graphs, maps, or pictures) introductory material which serves as the basis for the exercise. When writing, selecting, or revising introductory material, keep the material:
      (a) Relevant to learning objective(s) and the intellectual skill being assessed;
      (b) Appropriate for the examinees' development, knowledge and skill level, and academic or professional experience;
      (c) At a simple reading level, avoiding complex words or sentence structures, etc.;
      (d) Brief, as brief introductory material minimizes the influence of reading ability on testing;
      (e) Complete, contains all the information needed to answer items; and
      (f) Clear, concise, and focused on the IE's purpose.
   (2) Ensure that items, usually multiple choice:
      (a) Require application of the relevant intellectual skill (e.g., application, analysis, synthesis, and evaluation);
      (b) Don't require answers readily available from the introductory material or that can be answered correctly without the introductory material;
      (c) Are of sufficient number to be either proportional to or longer than the introductory material; and
      (d) Comply with item writing guidelines;
      (e) Revise content, as is necessary, when developing items.
   (3) When examinee correctly answers an item, it should be due to the fact that he or she has mastered the intellectual skill needed to correctly answer rather than failure to memorize background information. For

example, in statistics, give the examinee the formulas so that concepts are tested and not formula memorization.

(4) Other item formats are often included in the interpretative exercise.

C. Writing True/False Items

1. True/false items are commonly used to assess whether or not an examinee can identify a correct or incorrect statement of fact. Tests employing true/false items should contain enough items so that the relevant domain is adequately sampled and that the examiner can draw sound conclusions about examinee knowledge. True and false statements should be equally represented.

2. Strengths and limitations of true/false items (Gronlund 1998, p. 79; Oosterhof, 1994, pp. 155-158; & Popham, 2000, pp. 232-233) are:
   a. Strengths
      (1) These items enable an adequate sampling from the content domain.
      (2) They are relatively easy to construct.
      (3) They can be efficiently and objectively scored.
      (4) Reading ability exerts less of an influence than in multiple choice items.
   b. Limitations
      (1) These items are susceptible to guessing. There is a 50% chance of an examinee correctly guessing the correct answer given a true/false item.
      (2) Such items are only able to measure higher order skills indirectly.
      (3) These items can only be employed when dichotomous alternatives sufficiently represent reasonable responses.
      (4) Use of true/false items tends to encourage examinees to memorize facts, etc. as opposed to learning the content. If the item is not carefully constructed, student memorization "pays off."
      (5) True/false items provide no item diagnostic information as is found in multiple choice items by selecting the wrong answer.

3. True/false item writing guidelines (Gronlund 1998, pp. 80-84; Oosterhof, 1994, pp. 156-164; & Popham, 2000, pp. 235-238) are:
   a. For each statement, incorporate only one theme. When preparing an item, write it as both correct and incorrect. Select the better of the two options.
   b. Write crisp, clear, and grammatically correct statements. Brief is recommended. However, use terms that would suggest an incorrect response upon a quick reading by an examinee.
   c. Select precise wording for the item so that it is not ambiguous. The item should be either absolutely true or false; avoid statements that are partly true and partly false. If there are adverb or adjectives which are key to marking a statement correctly, underline or otherwise stress that word.
   d. Avoid double negatives. Negatively worded statements tend to confuse examinees so use them rarely.
   e. Attribute opinion statements to their source. However, if testing examinees' ability to distinguish fact from opinion, don't attribute.

f. If measuring cause and effect, use true statements.
g. Avoid absolute relative terms (e.g., always or never) usually adjective or adverbs, which tend to clue the examinees to the correct response.

4. True/False Item Variations
   a. True/false Items Requiring Corrections
      (1) This variation requires the examinee to either correct a false item or to identify the false portion of an item.
      (2) This is actually a blend of true/false and short answer.
      (3) **Directions**. Read the statement carefully. Determine whether or not each underlined number is true or false. If you think the statement is true, circle "T" or if you think the statement is false, circle "F." If false, place the correct number in the line provided.

      Acceptable reliability standards for measures exist. For instruments where groups are concerned, 0.80[a] or higher is adequate. For decisions about individuals, 0.85[b] is the bare minimum 0.95[c] is the desired standard.

      a. T F _____
      b. T F _____
      c. T F _____

      Answer: a. T; b. F 0.90; c. T

   b. Embedded Items
      (1) In a paragraph are included underlined words or word groupings. Examinees are asked to determine whether or not the underlined content possesses a specified quality, e.g., being true, correct, etc.
      (2) Embedded items are useful for assessing declarative or procedural knowledge.
      (3) **Directions**. Indicate whether the underlined word is correct within the context of threats to a measure's reliability. Circle "1" for correct or "2" for incorrect.

      When a test is given to a very similar homogeneous[a] (1 or 2) group, the resulting scores are closely clustered and the reliability coefficient will be high[b] (1 or 2).

      Answer: a. 1 and b. 2.

   c. Multiple True-False Items
      (1) This variation is a blend of the true/false and multiple choice item. Multiple statements, whose truth or falsity is being tested share a common stem. Each statement is numbered as a unique test item.
      (2) True/false item writing rules apply, but note that statements sharing a common stem are usually narrower in focus than conventional true/false items.

(3) **Directions.** Read each option and if you think the option is true, circle "T" or if you think the statement is false, circle "F."

The standard error of measurement is useful for
T F 1. Reporting an individual's scores within a band of the score range
T F 2. Converting individual raw scores to percentile ranks
T F 3. Reporting a groups' average score
T F 4. Comparing group differences

Answers: 1.T; 2.F; 3.F; 4.F

d. Focused Alternative-Choice Items
(1) While conceptually similar to true/false items, examinees are required to select between two words or values which correctly complete a statement. The words or values must have opposite or reciprocal meanings.
(2) When compared to conventional true/false items, focused alternative-choice items typically produce more reliable test scores.
(3) **Directions**. Read each statement carefully. Circle the letter that represents the word or phrase which makes the statement correct.

A B 1. A measure's reliability coefficient will likely (a. Increase or b. Decrease) as the numbers of items are lengthened.
A B 2. An examinee's answer to a restricted response essay should typically not exceed (a. 150 or b. 250) words.

Answers: 1. a; 2.a

e. Standard Format
(1) Traditionally, a true/false item has been written as a simple declarative sentence which was either correct or incorrect.
(2) **Directions**. Read the statement carefully. If you think the statement is true, circle "T" or if you think the statement is false, circle "F."

T F A measure can be reliable without being valid.

Answer: T

D. Writing Matching Items

1. The matching item is a variation of the multiple choice item format. Examinees are directed to match (i.e., associate) items from two lists. Traditionally, the premises (the portion of the item for which a match is sought) is listed on the left hand side of the page. On the right hand side of the page are listed the responses (the portion of the item which supplies the associated elements).
   a. The key to using this item format is the logical relationship between the elements to be associated.
   b. Elements in each column are homogeneous, i.e., related to the same topic or content domain.

c. There are usually more responses (right hand column) than premises. This reduces the biasing impact of guessing on the test score.
d. Example

Directions. Match the terms presented in the right column to their definitions which are presented in the left column. Write the letter which represents your choice of definition in the blank provided.

Definition	**Term**
_____ 1. Measures essential minimums that all examinees should know	a. Mastery Items
_____ 2. Designed to assess higher level concepts & skills	b. Power Items
_____ 3. Designed to measure what typical examinees are expected to know	c. Speed Items
_____4. Designed to measure achievement where the content is evolving	d. Independent Variable Items
	e. Dependent Items

Answers: 1.a; 2.b; 3.b; 4.d

2. Strengths and limitations of matching items (Gronlund, 1998, p. 85-86; Kubiszyn & Borich, (1996, p. 99; Oosterhof, 1994, p. 148; & Popham, 2000, pp. 238-240) are:
   a. <u>Strengths</u>
      (1) An efficient and effective format for testing examinee's knowledge of basic factual associations within a content domain.
      (2) Fairly easy to construct. Scoring is fast, objective, and reliable.
   b. <u>Limitations</u>
      (1) This format should not be used to assess other cogitative skills beyond mere factual associations which emphasize memorization.
      (2) Sets of either premises or responses are susceptible to clues which help under-prepared examinees guess correctly which gives rise to false inferences respecting examinee content mastery.

3. Item writing guidelines (Gronlund, 1998, p. 86-87 & Popham, 2000, pp. 240-241) are:
   (a) Employ only homogeneous content for a set of premises and responses.
   (b) Keep each list reasonably short, but ensure that each is complete, given the table of specifications. Seven premises and 10-12 options should be the maximum for each set of matching items.
   (c) To reduce the impact of guessing, list more responses than premises and let a response be used more than once. This helps prevent examinees from gaining points through the process of elimination.
   (d) Put responses in alphabetical or numerical order.
   (e) Don't break matching items across pages.
   (f) Give full directions which include the logical basis for matching and indicate how often a response may be used.

(g) Follow applicable grammatical and syntax rules.
(h) Keep responses as short as possible by using key precise words.

E. Writing Completion and Short Answer Items

1. Completion and short answer items require the examinee to supply a word or short phrase response. For example:

   Directions. Read each statement carefully. Write in the word or words which completes and makes the statement correct.

   1. The two most essential attributes of a measure are: (a) _____ and (b)_____.

   2. If a test is to be re-administered to the same examinees, the researcher is concerned about _____ reliability.

   3. If the reliability of a test is zero, its predictive validity will be _____.

   Answers: 1a. validity; 1b. reliability; 2. test-retest; 3. zero

2. Strengths and limitations (Gronlund, 1998, pp. 96-97; Kubiszyn & Borich, 1996, p. 99-100; Oosterhof, 1994, pp. 98-100; Popham, 2000, pp. 264-266) of completion and short answer items identified are:
   a. Strengths
      (1) These item formats are efficient in that due to ease of writing and answering, more items can be constructed and used. Hence, much content can be assessed, improving content validity.
      (2) The effect of guessing is reduced as the answer must be supplied.
      (3) These item formats are ideal for assessing mastery of content where computations are required as well as other knowledge outcomes.
   b. Limitations
      (1) Phasing statements or questions which are sufficiently focused to elicit a single correct response is difficult. There are often more than one "correct" answer depending on the degree of item specificity.
      (2) Scoring maybe influenced by an examinee's writing and spelling ability. Scoring can be time consuming and repetitious, thereby introducing scoring errors.
      (3) This item format is best used with lower level cogitative skills, e.g., knowledge or comprehension.
      (4) The level of technological support for scoring completion and short answer items is very limited as compared to selected response items (i.e., matching, multiple choice, and true/false).

3. Item writing guidelines (Gronlund, 1998, pp. 87-100; Oosterhof, 1994, pp. 101-104; & Popham, 2000, pp. 264-266) are:
   a. Focus the statement or question so that there is only one concise answer of one or two words or phrases. Use precise statement or question wording to avoid ambiguous items.
   b. A complete question is recommended over an incomplete statement. You should use one, unbroken blank of sufficient length for the answer.

c. If an incomplete statement is used, some authors recommend a separate blank of equal size for each missing word, others consider such an action to be a clue to examinees as to elements of the correct response. Avoid using broken lines which correlate to the number of letters in each word of the desired answer.
d. Place the blank at the end of the statement or question.
e. For incomplete statements, select a key word or words as the missing element(s) of the statement. Use this item variation sparingly.
f. Ensure that extraneous clues are avoided due to grammatical structure, e.g., articles such as "a" or "an".
g. Ensure that each item format allows for efficient and reliable scoring.

III. Constructing Restricted and Extended Response Items to Assess Analysis, Synthesis, and Evaluation

A. Introduction

1. A restricted response essay poses a stimulus in the form of a problem, question, scenario, etc. where the examinee is required to recall, organize, and present specific information and usually construct a defensible answer or conclusion in a prescribed manner.
   a. Restricted responses are best suited for achievement tests and the assessment of lower order intellectual (knowledge, comprehension, and application) skills.
   b. The restricted response essay provides the examinee with guidance for constructing a response.
   c. It also provides information on how the item is scored.
   d. Oosterhof (1994, p. 113) suggests that it should take examinees 10 or fewer minutes to answer a restricted response essay item. Limit the number of words in a response to a maximum of 100 or 150.
   e. Examples
      (1) Compare and contrast each of the four types of reliability. (Restricted Response)
      (2) Define each type of validity. (Restricted Response)

2. An extended response essay is one where the examinee determines the complexity and response length.
   a. An extended response essay is an effective device for assessing higher order cognitive (e.g., Bloom, et al.'s analysis, synthesis or evaluation) skills.
   b. This item format is significantly influenced by writing ability which may mask achievement deficits.
   c. Oosterhof (1994, p. 113) recommends that extended response essays not be used on tests given the limitations cited below. He does suggest that extended response items are useful as assignments.
   d. Examples
      (1) Illustrate how test validity and reliability using an original example involving at least one type of validity and two types of reliability would be established. (Extended Response)

(2) Using an original example, explain the process for conducting an item analysis study, including (a) a brief scenario description, (b) at least three sample items in different formats and (c) the computation, interpretation, and application of relevant statistical indices. (Extended Response)

3. Strengths and limitations of essay items (Gronlund, 1998, p. 103; Kubiszyn & Borich, 1996, pp. 109-110; Oosterhof, 1994, pp. 110-112; Popham, 2000, pp. 266-268) are:
   a. Strengths
      (1) These formats are effective devices for assessing lower and higher order cognitive skills. The intellectual skill(s) must be identified and require the construction of a product that is an expected outcome of the exercise of the specified intellectual skill or skills.
      (2) Item writing time is reduced as compared to select response items, but care must be taken to construct highly focused items which assess the examinee's mastery of relevant performance standards.
      (3) In addition to assessing achievement, essay formats also assess an examinee's writing, grammatical, and vocabulary skills.
      (4) It is almost impossible for an examinee to guess a correct response.
   b. Limitations
      (1) Since answering essay format items consumes a significant amount of examinee time, other relevant performance objectives or standards may not be assessed.
      (2) Scores may be influenced by writing skills, bluffing, poor penmanship, inadequate grammatical skills, misspelling, etc.
      (3) Scoring takes a great amount of time and tends to be unreliable given the subjective nature of scoring.

B. Writing and Scoring Restricted and Extended Response Essays

1. Item writing guidelines (Gronlund, 1998, pp. 87-100; Kubiszyn & Borich, 1996, pp. 107-109; Oosterhof, 1994, pp. 101-104; Popham, 2000, pp. 264-266) are:
   a. Use extended essay responses to assess mastery of higher order cognitive skills. Use restricted responses for lower order cognitive skills. Use Bloom, et al.'s Taxonomy or similar cognitive intellectual skills classification system to identify what cognitive skills are to be assessed.
   b. Use the appropriate verbs in preparing the directions, statement, or question to which examinees are to respond.
   c. Ensure that examinees have the necessary enabling content and skills to construct a correct response to the prompt.
   d. Frame the statement or question so that the examinee's task(s) is explicitly explained.
   e. Frame the statement or question so that it requires a response to a performance objective or standard.
   f. Avoid giving examinee's a choice of essays to answer. They will most likely pick the ones they know the answer to, thus adversely affecting content validity.

g. Have sufficient time to answer each essay item, if used on a test, and identify the point value. It's best to use restricted response essays for examinations.
h. Verify an item's quality by writing a model answer. This should identify any ambiguity.

2. Item scoring guidelines (Gronlund, 1998, pp. 107-110; Kubiszyn & Borich, 1996, pp.110-115; Oosterhof, 1994, pp. 113-118; Popham, 2000, pp. 270-271) are:
   a. Regardless of essay type, evaluate an examinee's response in relation to the performance objective or standard.
      (1) Ensure that the item's reading level is appropriate to examinees.
      (2) Grade all examinees' responses to the same item before moving on to the next. If possible, evaluate essay responses anonymously, i.e., try not to know whose essay response you are evaluating.
      (3) For high stakes examinations (i.e., degree sanction tests), have at least two trained readers evaluate the essay response.
   b. For restricted/brief response essays, ensure that the model answer is reasonable to knowledgeable students and content experts, and use a point method based on a model response. Restricted essays are not recommended for assessing higher order intellectual skills.
   c. For an extended response essay, use a holistic or analytical scoring rubric with previously defined criteria and point weights. When evaluating higher order cognitive skills, generally evaluate extended essay responses in terms of
      (1) Content: Is the content of the response logically related to the performance objective or standard being assessed? Knowledge, comprehension, or application skills can be assessed depending on how the essay item was constructed.
      (2) Organization: Does the response have an introduction, body and conclusion?
      (3) Process: Is the recommendation, solution, or decision in the essay response adequately supported by reasoned argument and/or supporting evidence? This is the portion of the response where analytical, synthesis, or evaluation skills are demonstrated. Again, be very specific as to which skill is being assessed and that the essay item, itself, requires the examinee to construct a response consistent with the intellectual skill being assessed.
   d. When evaluating essays where a solution or conclusion is required, consider the following in constructing your scoring plan:
      (1) Accuracy or reasonableness: Will posited solution or conclusion work?
      (2) Completeness or internal consistency: To what extent does the solution or conclusion relate to the problem or stimulus?

(3) Originality or creativity: Is an original or creative solution advanced? While an original or creative solution might not be expected, when one is posited, credit should be given.

IV. Statistical Strategies for Improving Select Response Test Items (Item Analysis)

A. Item Analysis: Purpose

1. The purpose of analyzing item behavior is to select items which are best suited for the purpose of the test.
2. The purpose of testing is to differentiate between those who know the content and those who don't. This can be done by:
   a. Identifying those items answered correctly by knowledgeable examinees.
   b. Identifying those items answered incorrectly by less knowledgeable examinees.
3. There are two indices: item difficulty or p-value, Index of Discrimination (D).

B. Item Analysis: Indices

1. Item Difficulty (p-value)

a. Item difficulty is determined by the number of examinee's correctly endorsing (i.e., answering) a dichotomously scored item. The p-vale has also been called the "item mean." The p-value is expressed as a proportion, with a range from "0.00 to 1.00."

b. Item format affects p-values, particularly where the item format is able to guess a response. For example:

(1) 37 x 3 = _____ (This format presents little opportunity to correctly guess--no choices.)

(2) 37 x 3 = _____ (This format presents more opportunity to correctly guess.)

(a) 11.1	(c) 1111
(b) 111	(d) 11.11

c. Target p-values

From the test designer's perspective, to maximize total test variance (needed for high reliability coefficients), each item p-value should be at or close to 0.50. Some recommend a range between 0.40 and 0.60 (Crocker & Algina, 1986, pp. 311-313).

d. Oosterhof (1994, p. 182) recommend differing p-value targets based on item format. These are:

(1) True-false and 2 option multiple choice, 0.85.
(2) Three option multiple choice, 0.77.
(3) Four option multiple choice, 0.74.
(4) Five option multiple choice, 0.69.
(5) Short-answer and completion, 0.50.

e. p-values are rarely the primary criterion for item selection into a test and for tests designed to differentiate between those who know the content and those who don't; p-values should be of consistent, moderate difficulty.

2. **Item Discrimination Index (D)**
   a. Items with high discrimination ability are prized as they contribute to sorting examinee performance. Determining an item's discrimination power requires computing "D," the item discrimination index. "D" is an important index when selecting items for a test whose purpose is to sort examinees based on knowledge. "D" works best with multiple choice and matching items.

   b. Formula: $D = P_u - P_l$
      (1) Term Meanings
         (a) D = discrimination index
         (b) P_u = upper proportion of examinees who answered item correctly
         (c) P_l = lower proportion of examinees who answered item correctly

      (2) To determine P_u and/or P_l, select the upper 25-27% of examinees and the lower 25-27% of examinees. Determine the proportion in each group answering the item correctly.

      (3) For example: Item 1, $P_u = 0.80$ & $P_l = 0.30$, so 0.80 - 0.30 = 0.50. Thus, D = 0.50. Eighty percent answered Item 1 correctly in the high scoring group; whereas, 30% did so in the lower scoring group. See Interpretive Guidelines below.

   c. Properties of "D"
      (1) Ranges from -1.0 to 1.0
      (2) Positive values indicate the item favors the upper scoring group.
      (3) Negative values indicate the item favors the lower scoring group.
      (4) The closer "D" is to 0.00 or 1.0 the less likely the item is going to have a positive value.

   d. Interpretative Guidelines (Crocker & Algina, 1986, pp. 314-315)
      (1) $D \geq .40$, well-functioning item—keep as is.
      (2) $0.30 < D > 0.39$, little or no item revision needed.
      (3) $0.20 < D > 0.29$, marginal item, needs revision.
      (4) $D \leq 0.19$, eliminate or completely revise item.

3. **Distractor Analysis** (For multiple choice or matching items)
   a. Distractor analysis assesses how well each multiple choice item response option (also called foils or distractors) attracts examinee endorsement (i.e., selection as correct answer); all foils should attract an endorsement (i.e., be selected). Don't use with True/False items.

   b. To determine an acceptable level of distractor functioning, one considers:
      (1) A distractor not attracting any endorsement is poorly functioning.
      (2) A distractor attracting an incorrect endorsement from a high scoring student is fine (most likely an error), provided there are only a few.

(3) The most likely incorrect response should have the 2'nd highest endorsement level. If all incorrect foils are equally likely, then lower scoring examinee selection should be fairly consistent across the incorrect foils.
(4) Revise foils which high- and low-scoring examinees together select.

c. Many commercial item analysis software programs will only provide an "all examinees" distractor analysis. However, understanding how distractors function is best achieved by reviewing high and low scoring student foil endorsements along with the "all examinees" option. Use the "high" and "low" scoring examinee groups used in an item discrimination analysis.

4. **Interpreting an Item Analysis Report**

a.

Item 6	A	B*	C	D	E	Omit
High	.07	.86	.00	.07	.00	.00 (Must = 1.0 or 100%)
Lower	.13	.60	.07	.13	.07	.00 (Must = 1.0 or 100%)
All	.12	.70	.05	.10	.03	.00 (Must = 1.0 or 100%)

p-value = 0.70 D = 0.26 * = Correct foil Rows = 1.0 or 100%

(1) Distractors function adequately. All foils attracted some endorsement.
(2) For a five distractor item, the p-value is okay, as the target is 0.70.
(3) The "D" value suggests some revision as it falls into the 0.20 to 0.29 range. Make foils "C," "D," and/or "E" more attractive to higher scoring students.

b.

Item 13	A	B	C*	D	E	Omit
High	.00	.00	100	.00	.00	.00
Lower	.00	.00	100	.00	.00	.00
All	.00	.00	100	.00	.00	.00

p-value = 1.0 D = 0.00 * = Correct foil Rows = 1.0 or 100%

(1) Unless the item is an "ice breaker," it is a very poorly functioning item. It doesn't meet any targeted p-value and has no discrimination ability. The item will not contribute to test score variance (i.e., reliability.).
(2) Item should be discarded or completely revised, unless it is an "ice breaker."

c.

Item 29	A	B*	C	D	Omit
High	.36	.36	.00	.28	.00
Lower	.41	.24	.10	.25	.00
All	.24	.60	.10	.06	.00

p-value = 0.60 D = 0.11 * = Correct foil Rows = 1.0 or 100%

(1) The target p-value (0.74) for a 4-foil multiple choice item suggests that the item has promise, but requires revision. However, the 0.11 "D" value indicates that the item has very little discriminating power.
(2) Distractors "A", "B", and "D", attracted similar endorsements from both the higher and lower scoring groups. This suggests item ambiguity.
(3) A revision strategy would be to make foil "C" more attractive and revise foils "A", "B", and "D" so that they are more attractive to lower scoring or higher scoring examinees, but not both.

5. It should be clear that improving test items, regardless of format, requires a balance between targeted p-values and acceptable "D" values. Revision requires a thorough understanding of the tested content and the cognitive (i.e., intellectual and thinking) skills required to correctly answer an item. The item analysis study model advanced by Crocker and Algina (1986, pp. 321-328) is instructive in guiding test developers to efficiently and effectively revise test items. The elements of the model are:
   a. Decide whether you want a norm-referenced or criterion-referenced score interpretation (see Appendix 5.1).
      (1) Do you want a norm-referenced (NRT) or criterion-referenced test (CRT) score interpretation? In either case, the IA indices discussed above are appropriate.
      (2) Given how content validity is established, it can be argued that test results (e.g., scores) can be used to assess instructional efficacy and as a basis for initiating remedial instruction, as NRT and CRT tests are constructed in much the same manner.
   b. Select relevant item performance indices.
      (1) For most IA studies, the parameters of interest are p-value, "D" and distractor analysis (for multiple choice items).
      (2) Comparing higher and lower scoring examinees is standard practice.
   c. Pilot the test items with a representative examinee sample for whom the test is intended.
      (1) Pilot test group size:
         (a) 200 examinees for local studies or 5-10 times the number of examinees compared to the number of items; for example, if there are 30 test items, a minimum of 150 examinees is recommend.
         (b) Several 1000 examinees are needed for large scale studies.
      (2) The above recommendations are for school district-wide, regional, or state-wide testing programs. It is not expected that classroom teachers/trainers developing examinations would need a pilot test with such large samples, but having a knowledgeable colleague(s) review your test is recommended.
   d. Compute each item performance indices (See Part IV of this Chapter.)

e. Select well performing items and/or revise poorly functioning ones.
   (1) Use target p-values and the item discrimination index criteria.
   (2) Crocker and Algina (p. 323) have described and resolved two common situations:
      (a) When there are more items than can be administered, the task is to select those items which make the most significant contribution to the desired level of test reliability and validity. Items with p-values between 0.40 and 0.60 are recommended as such items have been shown to reliably discriminate across virtually all ability groups.
      (b) If the item pool is not overly large and the test developer wants to keep every possible item, revise those items with low "D" values.
   (3) If necessary, repeat the pilot test with revised or newly written items.

f. Select the final set of items.
   (1) Use the "D" value and p-value criteria to guide decisions for test item selection and revising flawed items. Eliminate only as a last resort.
   (2) Look for unexpected or abnormal response patterns, particularly in distractor analysis.

g. Conduct a cross-validation study to determine whether or not the desired results have been attained.
   (1) This not usually done in classroom achievement testing. However, for reasonably high stakes tests such as a common departmental course final, one should consider conducting a validation study. Such a study might involve using the same or similar items several times and then taking the best functioning items for a final version of the examination.
   (2) At this stage, the test developer's interest is to see how well items function a second time. Some items will likely be removed. However, the examination is for the most part, constructed.

V. Constructing Direct Performance Assessments

A. Introduction

1. Direct performance assessments (DPA's) are used when specific processes (e.g., company procedures), skills (i.e., behaviors), outcomes (e.g., products), affective dispositions (e.g., attitudes, opinions, intentions), or social skills (e.g., self-direction, capability to work with others, manners, etc.) are to be directly assessed.

2. In direct performance assessment (DPA), examinees are presented a simulated or "real world" task, problem, or process where they are to demonstrate a procedure or construct a verbal, written, product, or blended product usually in the form of a demonstration, presentation, project, report, or term paper. The constructed response is then rated by a judge(s) (i.e., trainer, judge, teacher, professor, etc.), using checklists, rating scales or scoring rubrics.

a. Direct performance assessment inferences are highly contextually dependent. Any performance inference is based only on the demonstration or product produced and the context within which it was constructed or demonstrated.
b. Common assessment devices include checklists, rating scales, portfolios, and task descriptions with scoring rubrics.

3. Direct performance assessment can also be used to supplement traditional forms of assessment, assess higher order intellectual skills, and/or assess psychomotor skills.
   a. Traditional forms of testing (e.g., paper and pencil or computer based testing) will not allow inferences or judgments, based on direct performance observation.
      (1) Recall that traditional testing is most efficient, valid, and reliable for testing lower order intellectual skills (i.e., knowledge, comprehension, and simple applications).
      (2) Disposition (i.e., attitudes) assessment is most often indirect and employs rating scales (also called an index).
         (a) An examinee can individually respond to a rating scale to self-report his or her motivation, attitude, opinion, or intention.
         (b) Direct performance assessments, using behaviorally oriented rating scales, made by observation, can be used to assess or confirm indirect measurement of examinee motivation, intention, cooperation, social interaction skills, etc.
   b. Higher order intellectual skills (i.e., complex application, analysis, synthesis, and evaluation) can be assessed using DPA's.
      (1) Assessing higher order intellectual skills would include tasks which require the acquisition, organization, and use or application of information to real or simulated problems, scenarios, or opportunities.
      (2) Assessment devices or tools for these products include charts, tables, map reading, drawings, essay composition, experiments, projects, completion of tasks, etc.
   c. The assessment of psychomotor skills or the integration of intellectual and psychomotor skills includes tasks which require the examinee to provide an observable performance, which might take form as a speech, written communication, typing, dance, gymnastic routine, sales presentation, mock interview, or correctly cooking a recipe.

B. Direct Performance Assessment: Advantages and Limitations

1. DPA Advantages
   a. DPA usage allows for performance inferences or judgments which are impossible or at least very difficult to make with traditional testing strategies. Skill performance diagnosis and correction is also possible.

b. Higher order intellectual skills, built on lower order intellectual skills, can be directly assessed or tested through performance assessment. Well designed and implemented tasks can be "correctly" solved in a variety of ways thus facilitating examinee critical thinking skills and for the making of evaluative inferences by examiners.
c. Very clear connections to teaching or training quality can be made. Both the process of constructing the response (i.e., product) and the response itself can be assessed and evaluated; thereby improving teaching and learning, and assessment.
d. Examinees' commitment is increased given their sense of control over the learning, assessment and evaluation processes. Motivation tends to be high when examinees are required to produce an original work or product.

2. DPA Limitations
   a. Performance assessments are very labor and time intensive to design, prepare, organize, and evaluate. Records must also be maintained.
   b. Performance has to be scored immediately if a process is being assessed.
   c. Scoring is susceptible to rater error. Raters or examiners must be highly trained on the task description, performance criteria, and rating form to assess performance similarly and consistently. A plan for "breaking ties" is needed; two examiners may rate a performance quite differently. A third more senior or experienced rater can serve as the "tie breaker."
   d. Complex intellectual or psychomotor skills are composed of several different but complimentary enabling skills (e.g., reading level, drawing talent, physical coordination, stress tolerance, etc.) which might not be recognized or assessed. Examinees will likely perform some enabling skills more proficiently than others. Critical enabling skills should be identified and specifically observed and rated.
   e. Other potential limitations include time, cost, and the availability of equipment and judges. Due to these issues, performance assessment must be used to assess very highly relevant skills, which are teachable.

C. Constructing a DPA: The 3 Phase Process

1. Plan the Direct Performance Assessment (DPA).
   a. Senior examiners are responsible to determine what process, skill, product, or disposition is/are to be assessed; salient indicators which define realistic performance levels; the task specification; the scoring mechanics; and rater qualification and proficiency training.
      (1) Only key indicators or "primary performance traits" should be assessed so that the raters don't become overwhelmed, ineffectual, and inaccurate by being mired in excessive detail.
      (2) Ensure and document that examinees have mastered prerequisite knowledge and skills prior to launching a performance assessment.
      (3) Performance raters' or examiners' training is planned
   b. The outcome of phase 1 is a comprehensive DPA construction plan.

2. Construct the Individual DPA Elements (Stiggins, 1997)
   a. First, specify what decision (mastery, rank order, or a combination) is to be made. For example (Appendix 5.2), graduate students in an assessment course must successfully construct a direct performance assessment task according to performance criteria prepared by the professor at least to the "proficient level;" this is a mastery decision.
   b. Second, write the task description.
      (1) Is the exercise structured and/or natural? Natural might include a school or work environment which is familiar to the examinee.
      (2) What actual examinee performance is to be assessed in terms of content and skill foci, processes to be employed, and the work product to be produced?
      (3) What are the performance dimensions/criteria? These should have been documented in the DPA planning phase. For example (Appendix 5.2), graduate students' skill (working in pairs) in constructing a direct performance assessment task is measured across the dimensions of "task description clarity," "performance criteria," "scoring system," and "authenticity;" see the scoring rubric (labeled PTQAI).
      (4) Ensure the task description is understandable to intended examinees.
   c. Third, write the performance scoring mechanism.
      (1) Explicitly set and define performance levels. For example (Appendix 5.2), a project is assessed by the professor using a rating scale with four performance criteria ("Task Description," Performance Criteria," Scoring System," and "Authenticity") and three performance gradients or levels ("Meets Standard," "Mostly Meets Standard, etc.). The performance categories must represent realistic levels of examinee performance.
      (2) The scoring mechanism is constructed with points allocated to ensure consistent and valid performance segmentation. For example, see Appendix 5.2. Each work team's DPA will be reworked until performance criteria are minimally met at the "Proficient" level according to the professor's judgment.
         (a) Develop a draft of the checklist or rating scale, if the decision has been made to use either, see Part V.D below. Be sure to closely edit and proof.
         (b) Decide whether a holistic or analytic scoring strategy will be used as well as the type of scoring rubric (i.e., task-specific, skill-focused, or generic). See the more detailed discussion in Part V.E, below. A holistic rubric produces a single score across multiple performance criteria; whereas an analytical rubric produces scores for specific criteria, but the individual criterion scores are combined into one composite (i.e., total) score. An analytical rubric holds much instructional value.
   d. Fourth, the rater's training is designed; identify raters (i.e., examiners).
   e. The outcomes of phase 2 are the specific decision to be made, a task description, a performance scoring strategy, and a rater training schema.

3. Assemble and Field-test the Task Description and Scoring Plan
   a. Have qualified colleagues review, comment, edit and proof the task description and checklist, rating scale or scoring rubric. Ensure the two align.
   b. Recruit desired raters and thoroughly train through role-playing.
   c. Repeat pilot testing until satisfied with checklist, rating scale, or rubric performance.
      (1) Look carefully for administration difficulties, unclear items (i.e., indicators), incomplete decision-making information, accuracy of performance descriptions, etc.
      (2) Compute and interpret appropriate reliability indices, if the checklist, rating scale, or rubric format permits such computations; it should.
      (3) Evaluate the results; change evaluative criteria or standards, if needed.
   d. Prepare documents to inform examinees. Examinees should be told what is expected of them and be given copies of the task description, rating tool (checklist, rating scale, or rubric), and descriptions or definitions of the performance levels prior to instruction/training and "grading."

D. DPA Checklists & Rating Scales

1. Checklists
   a. A checklist is a list of specific, discrete actions to be performed by examinees which are to be observed by a rater. Often, the actions are listed in the expected order of performance.
      (1) Checklists typically employ a binary choice (e.g., "Met Expectation" or "Not Meet Expectations") with a "Did Not Observe" option.
      (2) Checklists are well suited to score procedures; but, is time consuming to construct, fairly efficient to score, and highly reliable and defensible, provided they are content valid.
      (3) Checklists provide examinees with quality feedback on performance. See Chapter 4 for more detailed information on writing checklists.
   b. If you have ever taken a first aid or CPR class, at least a portion of your performance was most likely assessed with a checklist. If you've not taken a first aid or CPR class, you should, if you can.

2. Rating Scales
   a. Rating scales are efficient in assessing dispositions (attitudes), work products, and social skills. Rating scales are more difficult to construct than checklists, but tend to be both reliable and defensible, provided they are content valid. Rating scales are efficient to score and can provide quality feedback to examinees.
   b. It is essential that each response option be fully defined and that the definition be logically related to the purpose of the rating scale and be progressive (i.e., represent plausible examinee performance levels). Like checklists, rating scales tend to be unidimensional in that each assesses one characteristic of performance. However, related unidimensional rating scales are often combined into multidimensional rating scales.

(1) See Appendix 5.6, Part A (presentation, Execution) for an example of a unidimensional rating scale. When Parts A, B (presentation Presence), and C (presentation, Technology) are combined, a multidimensional rating scale is produced. Performance levels are defined by key words, such as "Poor" or "Good."
(2) Presented in Appendix 3.1 is a unidimensional rating scale with five point response options. Appendix 5.5 is a dichotomously scored rating scale.

c. Rating scales tend to be analytical in that they are more descriptive and diagnostic than checklists. For example, problem solving is thought of as a series of sequential, but related steps and sub-steps. A rating scale will assess examinee performance across each key sub-step and step in the process. When using a checklist or rating scale, examinee performance should be compared only to the established performance scale and not other examinees.

3. Constructing and Scoring DPA Checklists & Rating Scales
   a. Construction Guidelines for DPA Checklists and Rating Scales
      (1) Ensure that the process, skill, product, disposition, or combination thereof, along with key indicators have been clearly and fully specified. Use only the number of indicators necessary to allow for import inferences to be accurately made. Rate only the key indicators.
      (2) Make sure that there are sufficient resources (e.g., time, money, equipment, and supportive colleagues) to successfully implement the performance assessment. Decide a priori whether or not (a) examinees are permitted reference materials, (b) examinees have sufficient enabling knowledge to successfully perform, and (c) what are the scoring criteria.
      (3) Estimate the amount of time for virtually all examinees to complete the task. If sufficient time is not available, then reframe the task.
      (4) Use the fewest, most accurate words possible to define the points on the checklist or rating scale performance scoring continuum. Avoid using redundant phrases. Telegraphic phrases, centered on key nouns and verbs, which are understood by each rater are recommended.
      (5) Ensure that the form, itself, is efficient to score. Rater responses should be recorded by circling a number or using an "X" or "Y" or "N" (see Appendix 5.4 or 5.5 are examples). For either checklists or rating scales, provide a convenient space to mark or circle. Provide space for brief comments.
      (6) When rating a process, individual items must be grouped in the correct sequence. For product assessments, items of similar content or stage of production should be grouped together.
      (7) Keep item polarity (i.e., response options consistent).
         (a) For rating scales, place the least desired characteristic or lowest rating (e.g., "Strongly Disagree") on the left end of the continuum

and the most desired quality or highest rating on the right end of the continuum (e.g., "Strongly Agree").

(b) Arrange similarly for checklists (e.g., "Poor Quality" "Acceptable Quality," or "High Quality").

(8) Provide a space for rater comments beside each item or at the end of the checklist or rating scale.

(9) Always train examiners or raters so they proficient in using the checklist or rating scale.

(a) After each practice session, debrief the raters so that all can understand each other's logic in rating an examinee's performance, and where necessary arrive at a rating consensus.

(b) Stress to raters the importance of using the full range of rating options on rating scales; many raters use only the upper tier of the rating continuum, which is fine provided all examinees actually perform at those levels.

b. Scoring DPA Checklists and Rating Scales

(1) The binary choice (e.g., "No" or "Yes") can be weighted "1" or "2," respectively (Appendix 5. 3) and then summed to arrive at a total score and if items are grouped, subtest scores as well. Avoid weighting the "Not Observed" Option. For most checklist applications, the "Not Observed" should never be endorsed.

(2) The multi-point rating options (see Appendix 4.6 for other examples), used in rating scales, can be weighted where "1" (Strongly Disagree) represents the lowest rating and "5" (Strongly Agree) the highest, if a five point scale is used. After point weighting, total and/or subtest scores can be computed. See Appendix 4.9 on writing scorable items.

(3) A common strategy is to place examinees into performance categories based on their ratings. For an example, see Appendices 3.1 and/or 4.9, Table 4.9.2.

E. Constructing DPA Scoring Rubrics

1. A scoring rubric, similar to a rating scale, is (a) composed of multiple rating subscales or subtests, which when assembled together, represent the complete process, skill, or product to be rated (e.g., Appendix 5.6) or (b) a series of performance descriptions (e.g., Doesn't, Meets, or Exceeds Expectations) which address multiple qualities simultaneously using a common scoring scale (Appendix 5.5).

2. Definition of Terms

a. A rubric is an instrument for rating a constructed examinee response to a stimulus or task description and is either holistic or analytical.

b. Evaluative criteria are performance standards which are rated, typically using points. See the 33 standards identified in Appendix 5.3. Each performance standard should be logically related but functionally independent of other standards. Group related evaluative criteria together; but in the proper ordered sequence.

c. Quality Definitions are various performance levels (e.g., "Exceeds Expectations," "Meets Expectations" or "Doesn't Meet Expectations") which are specified and defined so that differences in examinee performance are determined. Each performance level needs its own definition.

3. Scoring Strategy is either holistic or analytic.
   a. Holistic: All evaluative criteria are considered but summarized into a single, overall quality judgment, usually presented in a single score.
      (1) While fairly easy to adapt or write, holistic rubric reliability tends to be weak and generalizability very limited.
      (2) The holistic rubric is subject to scoring bias and rater inconsistencies.
      (3) Unless holistic rubrics are well defined and detailed, raters are likely to substitute their personal opinions or definitions. Each performance description must represent plausible examinee performance levels. These definitions should be written by senior examiners who are very familiar with the process, skill, or product being rated.
      (4) The holistic rubric is satisfactory (a) if an overall appraisal of a specific skill or ability can be achieved and makes sense and (b) if used on less critical process, skills, or products and not used when high stakes decisions are to be made about an examinee's performance.
      (5) Holistic Scoring Rubric for rating a business memo
          6 = The memo is exceptional in conforming to the model format provided and meeting writing mechanics and punctuation expectations.
          5 = The memo is very good in conforming to the model format provided and meeting writing mechanics and punctuation expectations.
          4 = The memo is a good in conforming to the model format provided and meeting writing mechanics and punctuation expectations.
          3 = The memo is adequate in conforming to the model format provided and meeting writing mechanics and punctuation expectations.
          2 = The memo is poor in conforming to the model format provided and meeting writing mechanics and punctuation expectations.
          1 = The memo is very poor in conforming to the model format provided and meeting writing mechanics and punctuation expectations.

   b. Analytic: Each evaluative criteria is scored and contributes directly to an examinee's subtest and/or total score.
      (1) If the focus of the rating exercise is to assist individual examinees to improve process or skill performance, then analytic rating is recommended, despite its time intensive nature and high cost.
      (2) Analytic rubrics are richer in detail and description, thereby offering greater diagnostic value. This provides examinees with a detailed assessment of strengths and weaknesses.

(3) Analytic rubrics should be based on a thorough analysis of the process, skill, or task to be completed. Performance definitions must clearly differentiate between performance levels. Detailed checklists and rating scales are often used in analytic ratings.
(4) Analytic rubrics are presented in Appendices 5.2, 5.4, and 5.5.

4. Types of Analytical Scoring Rubrics
   a. Task-specific rubrics (Appendix 5.3) focus on a particular task rather than the skill associated with that task. Task specific rubrics contribute to improving inter-rater reliability. Popham (2000, p. 290) recommends against use. However, Haladyna (1997, pp. 134) has observed:
      (1) Task-specific scoring rubrics are expensive to create; but, if well-constructed, it can be used repeatedly.
      (2) Task-specific rubrics are both valid and reliable and provide examinees with accurate performance information.

   b. Skill-focused rubrics (Appendix 5.5) center on the skill being assessed. Haladyna's comments on task-specific rubrics apply to skill-focused rubrics. Popham (2000, p. 290) recommends the use of skill-focused rubrics. He argues that skill mastery is more instructionally relevant than task mastery. Popham provides guidelines for constructing a skill focused rubric:
      (1) The rubric should consist of 3 –5 evaluative criteria.
      (2) A teachable attribute or sub-skill should form the basis of the evaluative criterion
      (3) The evaluative criteria should apply to any task designed to assess student skill mastery within the same or similar context.
      (4) The format of the rubric should be well organized.

   c. Generic rubrics are applicable across a variety of contexts, such as scoring a written assignment regardless of content, topic, or author. See Figure 5.1. The numbers in parenthesis are point values associated with each performance level; 20 points are possible (4 points x 5 criteria).
      (1) While generic rubrics improve the generalizability of results, Haladyna (1997, p. 134) has identified two problems with generic rubrics:
         (a) "[A] single rating scale seldom leads to a reliable test result"
         (b) "[N]o ability [skill or task] is so simple that a single rating scale can cover it."
      (2) Where possible, convert a generic rubric to either a task or skill focused rubric.

Criterion	Novice (1)	Apprentice (2)	Proficient (3)	Exemplary (4)
Content	Little or no substantive content was provided.	Content was partially substantive, but not meeting expectation.	Content was substantive, meeting expectation.	Content was substantive, exceeding expectation.
Organization	Organization was lacking.	Organization partially met expectation.	Organization met expectation.	Organization exceeded expectation.
Originality	The writer provided little or no insight into the topic.	The writer's insightfulness partially met expectation.	The writer's insightfulness met expectation.	The writer's insightfulness exceeded expectation.
Critical Thinking	The critical thinking shown was limited or not in evidence.	The critical thinking demonstrated partially met expectation.	The critical thinking demonstrated met expectation.	The critical thinking demonstrated exceeded expectation.
Writing Mechanics (Spelling, punctuation, etc.)	The writing mechanics were poorly executed or not in evidence.	The writing mechanics demonstrated partially met expectation.	The writing mechanics demonstrated met expectation.	The writing mechanics demonstrated exceeded expectations.

Figure 5.1 Generic Rubric Example

References

Airasian, P. W., (1997). *Classroom assessment* (3rd ed.). New York, NY: McGraw-Hill.

Bloom. B. S., Engelhart, M. D., Frost, E. J., & Krathwohl, D. (1956). *Taxonomy of educational objectives. Book 1 cogitative domain*. New York, NY: Longman.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, & Winston.

Gagne, R. M. (1985). *The conditions of learning and theory of instruction* (4th ed.). Chicago, IL: Holt, Rinehart, & Winston, Inc.

Gronlund, N. E. (1998). *Assessment of student achievement* (6th ed.). Needham Heights, MA: Allyn & Bacon.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Needham Heights, MA: Allyn & Bacon.

Kubiszyn, T. & Borich, G. (1996). *Educational testing and measurement*. New York, NY: Harper Collins College Publishers.

Lyman, H. B. (1998). *Test scores and what they mean* (6th ed.). Needham Heights, MA: Allyn & Bacon.

Mitchell, R. (1996). *Front-end alignment: Using standards to steer educational practice*. Washington, DC: The Education Trust.

Oosterhof, A. (1994). *Classroom applications of educational measurement* (2nd ed.). New York, NY: Merrill.

Popham, W. J. (2000). *Modern educational measurement* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Quellmalz, E. S. (1987). Developing reasoning skills. In Joan Baron & Robert Sternberg (eds.) *Teaching thinking skills: Theory and practice* (pp. 86-105). New York, NY: Freeman.

State of Florida (1996*). Florida curriculum framework: Language arts PreK-12 Sunshine State standards and instructional practice*. Tallahassee, FL: Florida Department of Education.

Stiggins, R. J. (1997). *Student-centered classroom instruction*. Columbus, OH: Merrill.

Stiggins, R. J., Griswold, M. M., & Wikelund, K. R. (1989). Measuring thinking skills through classroom assessment. *Journal of Educational Measurement, 26*, (3), 233-246.

Appendices

Appendix 5.1 outlines Bloom's Taxonomy of Intellectual Skills and selected specific thinking skills. Knowing what intellectual skill is required for examinees to correctly answer an item is critical in constructing that item. Higher-order intellectual skills (analysis, synthesis, and evaluation) require the use of specific thinking skills some of which are described. The intellectual skill application may also require the use of specific thinking skills.

Appendix 5.2 is an interpretive exercise example. Multiple choice items can be constructed such that they assess higher-order thinking skills. An interpretive exercise has a core prompt or scenario around which several multiple-choice items are constructed.

Appendix 5.3 is a task specific, analytical rubric which includes a detailed description of the task to be accomplished and the performance scoring rubric.

Appendix 5.4 is a simple rating scale that was used by team members to rate the contribution of each other's contribution in the construction of a class project.

Appendix 5.5 is a skill focused scoring rubric used to rate student critiques of an article from a juried journal. It has a very brief task description but most of the performance criteria are presented in the rubric. The score for each criterion is summed to reach a final performance rating on the assignment.

Appendix 5.6 is a three-dimensional (three subtest) skill focused, analytical scoring rubric measuring presentation performance.

Appendix 5.6 describes ethical test preparation strategies for examinees as well as guidance on test taking skills.

Appendix 5.1
Classroom Assessment: Introduction

In a "standards based" approach to education and training, informed by Constructivist ideology and motivated by high stakes accountability, assessment informed instruction is the expectation as is continuous improvement. Assessment informed instruction requires the educator (teacher, trainer, planner, instructional designer or administrator) to plan, deliver, and adjust instruction based on students' or trainees' evolving mastery of learning and skill standards until the desired mastery is achieved.

The Teaching/Assessment cycle is outlined in Figure Appendix 5.1.1. Based on learning standards, teaching is conducted. Once teaching is launched, continuous formative assessment is engaged as is re-teaching based on assessment results. The assessment/re-teaching cycle is repeated until suitable mastery is demonstrated via summative assessment. Then a new teaching/assessment cycle begins. The teaching/ assessment cycle assumes that instruction and assessment are planned and executed in conformance to specified learning and performance standards.

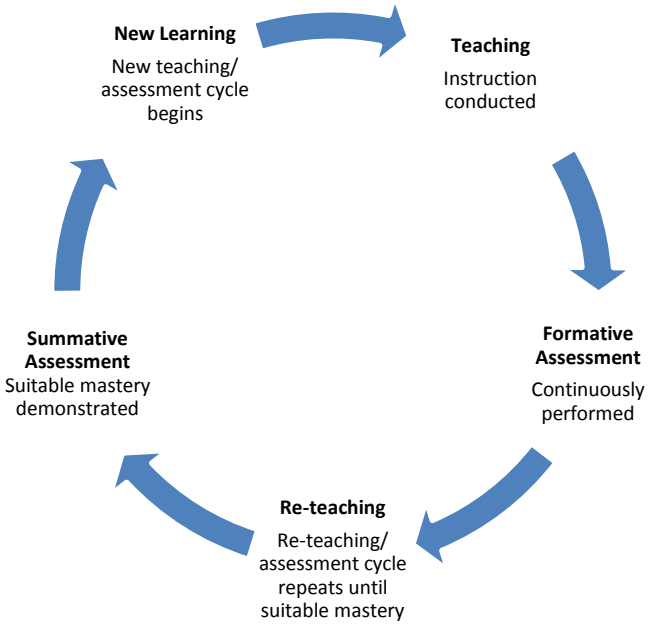

Figure Appendix 5.1.1 Teaching/Assessment

With the Teaching/Assessment Cycle as our application framework, in this chapter we will review the relationship between knowledge, learning, and intellectual skills. We will also examine the relationship between measurement, assessment, and evaluation.

I. Knowledge, Learning, Intellectual and Thinking Skills

A. Definition of Knowledge

1. Alexander (1996, p. 89) writes that knowledge "is a scaffold that supports the construction of all future learning." Greeno, Collins, & Resnick (1996, p. 16) argue that the cognitive view of knowledge "emphasizes understanding of concepts and theories in different subject matter domains [e.g., reading or science] and general cognitive abilities, such as reasoning, planning, solving problems, and comprehending language."
2. Knowledge can also be broadly categorized according to use, as declarative, procedural, or conditional (Paris & Cunningham, 1996; Paris, Lipson, & Wixson, 1993).
   (a) Farnham-Diggory (1994, p. 468) defined **declarative knowledge**, "knowledge that can be declared, usually in words, through lectures, books, writing, verbal exchange, Braille, sign language, mathematical notation, and so on." Declarative knowledge can be simple facts, generalities, rules, personal preferences, etc.
   (b) Woolfolk (2001, p. 242) defines **procedural knowledge** as "knowing how to do something such as divide fractions or clean a carburetor. Procedural knowledge must be demonstrated." Other examples of procedural knowledge include translating languages, classifying shapes, reading, or writing. In intellectual skills taxonomies proposed by Bloom, Engelhart, Frost, & Krathwohl (1956) and Gagne (1985), the levels beyond knowledge, are procedural knowledge.
   (c) Woolfolk (2001, p. 243) defines **conditional knowledge** as, "knowing when and why to apply…declarative and procedural knowledge." Conditional knowledge involves judgment. Examples of conditional knowledge include how to solve various math problems, when to skim or read for detail, when to change strategies when confronted with a perplexing problem, etc.

3. When measuring the effectiveness of instruction or an instructional program, the instructional designer or examiner may use traditional classroom testing strategies or direct performance assessments (Both are discussed in Chapter 5). Regardless of the strategy or mix of strategies selected, declarative knowledge, procedural knowledge, and/or conditional knowledge will be assessed. Thus, it is important to know which type of knowledge is being assessed in order to frame test items or construct direct performance assessments that will yield the information sought.
   (a) When examinee answers a multiple choice or true false item correctly, he or she displays declarative knowledge.
   (b) When an examinee answers an item which requires the listing of steps to bake cookies, according to a particular recipe, he or she displays procedural knowledge.
   (c) When an examinee is required to solve an algebraic expression, he or she displays conditional knowledge as he or she must know which mathematical or algebraic laws or procedures to apply to correctly solve

the problem. When the circumstances surrounding a business opportunity change, one must decide on whether to pursue the same strategy or change the strategy in order to win a business contract. Remember, conditional knowledge requires judgment and relies on declarative and/or procedural knowledge.

B. Learning

1. The process of acquiring knowledge (declarative, procedural, and/or conditional) is called learning. In order to assess instructional effectiveness, learning must be measured. Kimble defined learning as "[A] relatively permanent change in behavioral potentiality that occurs as a result of reinforced practice" 1961, p. 6). In other words, a learner must display his or her knowledge through behavior (e.g., answering a test item, repairing a car engine or modeling a particular attitude). Hergenhahn and Olson (1997, p. 2) have pointed out
   a. Leaning must be exhibited through behavior.
   b. Learning is a consequence of experience (e.g., life, schooling, training, practice, observation, etc.).
   c. Only reinforced (positively or negatively) experience, practice, etc. is learned. Reward is only one type of reinforcement.

2. Instructional design and teaching strive to provide experiences and reinforcement which enables one to learn, i.e., resulting in a permanent (or the realistic potential for) behavior change. It is this behavior that is measured and then based on that measurement, inferences are made about what has been learned, how well it has been learned, and how competently it may be applied. For example:
   a. The first grade student, who does not know how to read, learns to read. This is a permanent change in behavior. The teacher knows the student can read because the student read a story.
   b. The worker, who lost her job due to changes in technology, learns new job content and skills by going to a vocational-technical school. An employer can determine whether or not the applicant can repair small engines, by watching her diagnose and repair a broken lawnmower motor.
   c. A college freshman can demonstrate his knowledge of history by correctly answering several test items on causes of the Great Depression.

3. So one may ask, "How do we measure learning?"
   a. An instructional designer or teacher specifies the knowledge (declarative, procedural, and/or conditional); skills; or attitudes (KSA's) which need to be learned.
      (1) These KSA's are then expressed as learning targets, learning outcomes, learning standards, or learning objectives. (These terms mean the same thing.)

(2) Next, a learning target is "broken down" into its component parts (often called benchmarks) which a learner must know in order to achieve or master the learning target. So, when a learner is able to accomplish all benchmarks, we infer that he or she has achieved or mastered the learning target. See chapter 5 for more detail.

b. Once the learning targets and benchmarks are written, instructional materials are identified and sorted into modules or units. This sorting is an iterative process that often leads to changing learning targets, benchmarks, and/or instructional materials so that a "cleaner" alignment is achieved.

c. Specific instructional strategies are devised to facilitate learning.

d. Since knowledge must be measured through learner behavior (e.g., answering test items, writing a paper, or producing a work product), formative and summative assessments are devised and administered to learners in order to measure teaching or instructional design effectiveness and student learning.

(1) Based on formative assessment results, instruction may be adjusted to assist those who are not learning as intended or to accelerate learning if learners exhibit mastery faster than anticipated.

(2) At the conclusion of the learning experience, learners are usually given a summative assessment (e.g., test) to measure their learning. From these summative results, inferences are made about the effectiveness of the instructional design of the curriculum, instruction, and learning.

4. The most critical component of a formative or summative assessment are the test items to which students or examinees must respond. Test items are written to match learning target benchmarks. If the examinee achieved (or mastered) the learning target benchmark, he or she achieved the benchmark.

a. A test or instructional designer must be able to identify the mental (i.e., intellectual skills) a learner must possess in order to meet the learning target. Bloom and colleagues (1956) have has classified "knowledge" into six intellectual skills: knowledge, comprehension, application, analysis, synthesis, and evaluation; a detailed discussion follows. Each intellectual skill, written into a learning target benchmark, must be written into test items written to specifically measure or assess learner mastery of that benchmark.

b. A learning target benchmark may require an examinee to use a specific thinking skill to correctly answer its corresponding test item(s).

c. Test item writers must know knowledge type, the specific intellectual skill, and any specific thinking skill, an examinee must possess to answer the item correctly to show that the benchmark has been achieved.

C. Bloom, et al.'s Intellectual Skill Taxonomy

1. Bloom, et al.'s (1956) Taxonomy of Intellectual Skills

a. Bloom, et al. (1956, p. 201) defines **knowledge** (i.e., recalling or remembering) to include the recall of facts, methods, processes, patterns, structures, settings, etc.

(1) Knowledge is stored in the brain; the purpose of measurement is to present a stimulus (test item) which will clue the examinee to recall the stored knowledge. Kubiszyn and Borich (1996, p. 60) say knowledge is what students must remember. Learners must have this declarative knowledge as it is the basis for all higher order intellectual skills.
(2) When writing learning standards (i.e., targets, outcomes, or objectives) at the "Knowledge" level, use action verbs such as: list, name, recall, state, underline, write, record, count, recite, draw, find, match, choose, label, remember, recognize, select, define, list, etc.

b. **Comprehension** (i.e., really "getting it" or understanding) is the lowest of the higher order intellectual skills in the taxonomy; but learners use the declarative knowledge largely within the context in which it was learned.
(1) Examinees are expected to
(a) Translate knowledge from one form to another without losing its essential meaning within the original learning context;
(b) Interpret knowledge so as to identify its central elements or ideas, and then make inferences, generalizations, or summaries but within its original context or application; or
(c) Use the knowledge to extrapolate trends, implications, consequences, etc., within the original learning context.
(2) When writing learning standards at the "Comprehension" level use action verbs such as: compare, describe, restate, identify, contrast, express, explain, outline, paraphrase, summarize, report, convert, distinguish, estimate, infer, predict, rewrite, summarize, translate, etc.

c. **Application** is the use of knowledge in either an extension of the original learning situation or a new but related context. Procedural rules, technical principles, theories, etc. are examples of what must be remembered and applied. Application uses procedural or conditional knowledge.
(1) Examinees may be expected to:
(a) Translate knowledge from one form to another (verbal to written) without losing its essential meaning within a new/different context;
(b) Identify and Interpret the central elements or ideas of previously learned content/skills and then make inferences, or generalizations to a <u>new</u> context or situation (using conditional knowledge); or
(c) Apply procedural knowledge to solve a problem.
(2) Examples are: (a) diagnosing an automobile starter problem given prior experience with the same problem but with a different car; (b) answering a math word problem, using the different laws of addition, subtraction, and multiplication or solving equations; or (c) predicting a probable change in a dependent variable (grades) given a change in the independent variable (hours spent studying).
(3) <u>The key distinction between application and comprehension</u> is that examinees or students are required to perform what is or was "comprehended" in a "new" environment (e.g., situation).

(4) When writing learning standards at the "Application" level, use action verbs such as: apply, complete, demonstrate, interpret, illustrate, perform, operate, produce, role-play, distinguish, compute, construct, manipulate, modify, operate, predict, prepare, relate, show, solve, etc.

d. **Analysis** is breaking-down, "deconstructing," or "backwards engineering" a communication, theory, process, or "other whole" into its constituent elements or parts so that relationships (horizontal, vertical, diagonal, or hierarchical) between the component parts are made explicit.

(1) Analysis reveals the internal organization, assumptions, biases, of an argument (i.e., idea or position), theory, interpretation, problem, communication, process, or an opportunity, etc. Ask these questions:
   (a) Is the communication logically constructed?
   (b) Are the interpretation or argument's assumptions realistic?
   (c) Do the argument's component parts "fit" logically together?
   (d) Are there any logical fallacies?
   (e) Are there any biases shown?

(2) To respond to analysis level test items, the student or examinee might:
   (a) Deconstruct an <u>argument, theory, evidence, or interpretation</u> (expressed verbally or written) to recognize unstated assumptions, separate fact from conjecture, identify motives, separate a conclusion from its supporting evidence, and identify logical (or illogical) contradictions or inferences.
   (b) Once the constituent parts of a position or rationale for action (or inaction) have been identified, the relationships between those parts may be examined. It may be necessary to:
      [1] Revise or delete elements which are less critical or not strongly related to the position or rationale or
      [2] Analyze how the position or rationale was structured or organized, i.e., identify its organizing principles and techniques (e.g., form, pattern, etc.) to improve clarity or persuasiveness.
   (c) To analyze a potential <u>opportunity</u> (e.g., a potential strategic shift in a firm's direction), one might conduct an SWOT analysis, which means assessing the organization's strengths, weaknesses, opportunities, and threats, concerning the opportunity.
   (d) To analyze a <u>problem</u>, one must determine potential causes, how each potential cause operates or contributes to the problem and its consequences. Next, relationships between and among potential causes must be examined and understood before possible solutions can be crafted.

(3) When writing learning standards at the "Analysis" level use action verbs such as: compare and contrast, diagram, deduce, differentiate, show differences or similarities, analyze, critique, disassemble, distinguish or discriminate between, characterize, separate etc.

(4) Do not to confuse analysis with "Comprehension" or "Evaluation."
   (a) An examinee understanding or comprehending the content of a message, argument, interpretation, etc. displays "Comprehension."
   (b) "Evaluation" involves making a judgment about merit or worth, using explicit external criteria. See "Evaluation" below.
(5) Tasks or test items (i.e., brief or extended response items) built to assess "Analysis," will take time and perhaps even have more than one plausible correct answer or skill demonstration.

e. **Synthesis** is the production of a new or unique story, paper, article, book, poem, play, video, movie, argument, interpretation, plan, theory, or process etc. In synthesis new, similar, or different elements are combined into a new "whole."
   (1) When displaying the intellectual skill of "Synthesis", he or she may
      (a) Produce a unique original communication to inform an audience or reader about the author's ideas, feelings, experiences, etc. Influencing factors on these communications include its desired audience effect, nature of the audience, communication medium, conventions and forms of the medium selected to convey the communication, and the student or examinee him or herself. All of these elements combine to produce the new communication.
      (b) Produce a plan or proposed set of operations, i.e., a new or better procedure for doing or accomplishing something. The plan or procedure is the product, which must meet the requirements of the task, (e.g., product specifications).
      (c) Derive a new or different set of abstract relationships where the student or examinee constructs a new (to him or her) theory of human motivation, learning, leadership, or behavior, etc. Two examples are:
         [1] The student or examinee starts with concrete data or phenomena and must explain or classify the data or phenomena. Examples include the periodic table, biological phyla, developing taxonomy of intellectual skills, proposing a theory of personality or intelligence or hypothesizing a web of relationships between animals and plants in an ecosystem.
         [2] The student or examinee starts with basic propositions or hypotheses, and then deduces (i.e., or suggests) other propositions or relationships. The student must reason within a fixed framework. Examples include (a) formulating a theory; (b) testing a hypotheses using data from a study of the "best" way to teach math; or (c) modifying a theory or hypotheses based on newer or different data (i.e., new and improved "best" way to teach math).

(2) When writing learning standards at the "Synthesis" level use action verbs such as: construct, combine, compile, assemble, compose, formulate, design, revise or rewrite, organize, plan, prepare, propose, research, tell, generate, etc.
(3) "Synthesis" differs from "Comprehension," "Application," and "Analysis."
   a. "Synthesis" emphasizes creativity (uniqueness and originality) more than the other intellectual skills.
   b. Applying knowledge as taught in its original context is "Comprehension," in a changed, or entirely new context, its "Application."
   c. "Analysis" disassembles a "whole" for better understanding; "Synthesis" requires the examinee to assemble many different elements from different sources to construct a "new whole."
   d. Test items or tasks at the "Synthesis" level have more than one correct answer. Assignments or tasks that require the student to function at the "Synthesis" level enhance creativity, but a thorough knowledge of the content or skill domain is required.

f. **Evaluation** involves the application of evaluative criteria to procedures, processes, ideas, people, products, art works, solutions, etc. for the purpose of making a judgment about merit or worth.
(1) These judgments are based on internal or external evaluative criteria.
   (a) Evaluative Judgments using internal criteria focus on the accuracy of the work (e.g., term paper grading, novel, idea, problem solution, etc.). For example in a written work, attention is given to its internal logic, consistency, and lack of flaws. Indicators include the consistent use of terminology, flow, relationship of conclusions or hypotheses to the data or evidence presented, precision and exactness of words and phrases, reference citations, writing mechanics, etc. Considered together, the indicators influence perceptions of accuracy and quality.
   (b) Evaluative Judgments using external criteria are made about paintings, employee or athlete performance, a gymnastic routine, or work procedure, etc. Evaluative judgments must be made using criteria drawn from the relevant discipline, trade, or sport. For example, a work on nursing must be evaluated in terms of nursing criteria; art or literature in terms of the genre's governing conventions; or an assignment in light of its scoring rubric.
(2) When writing learning standards at the "Evaluation" level use action verbs such as: judge, assess, appraise, justify or defend, support, score (as in applying a rubric), conclude and support, prove and support, rank and support, select or recommend and explain, criticize, critique, defend, etc.

2. Operationalizing Bloom, et al.'s Taxonomy for Assessment
    a. For writing benchmarks, use the most precise action verb to enable a test or instructional designer to write test items or tasks (see Chapter 5) which require the examinee to demonstrate he or she possesses the specified intellectual or specific thinking skill (see below).
    b. An explicitly written benchmark, will enable required intellectual and specific thinking skills to be designed into instructional materials and learning experiences and ensure they are taught, learned, and tested. Benchmarks drive test item and task description writing.
    c. There has been a revision of Bloom's Taxonomy; see Anderson and Krathwohl (2001). Webb's Depth of Knowledge model is increasingly used ("*Webb's Depth of Knowledge Guide*," 2009; "*Depth of Knowledge Levels*," 2005). For a comparison of Bloom's Taxonomy and Webb's Depth of Knowledge see "*Levels of Thinking in Bloom's Taxonomy and Webb's Depth of Knowledge"* (n.d.). For a comparison of multiple intellectual skills taxonomies see Hess (2006).

D. Specific Thinking Skills

1. Introduction
    a. To design valid prompts or task descriptions for brief or extended response test items, term papers, or projects, an instructional designer or examiner must incorporate or teach the specific thinking skills required by the controlling benchmark(s) to ensure that the examinee can correctly answer the test item or complete the assigned task.
    b. The intellectual skills most associated with thinking skills are "Application," "Analysis," "Synthesis," and "Evaluation." For example, critical thinking requires the use of "Analysis," "Synthesis," and/or "Evaluation." Rarely, are thinking skills associated with "Knowledge" or "Comprehension;" these are most frequently used with select response items (e.g., multiple-choice, fill in the blank, matching, or true/false).
    c. Frequently, a project, task, or test item requires examinees to use multiple thinking skills. Additionally, the purpose of instruction may be to teach or develop an intellectual or thinking skill. In either case, care ensure that examinees possess these skills and can effectively use them.
    d. There are multiple thinking skills taxonomies, e.g., Kagan (2003) or Perkins (as cited in Brandt, 1986), or Sale (n.d.). Identify the thinking skill taxonomy most closely aligned with your work discipline; become thoroughly familiar with it.

2. Specific Thinking Skills
    a. Creative thinking is closely associated with writing fiction, plays, short stories or creating new and innovative procedures or processes. Michalko (1991, 2001) offers specific strategies to improve creative thinking. De Bono (1999) provides differing perspectives on creative thinking.

b. Critical thinking will typically involve the use of analysis, synthesis, and evaluation in order to carefully critique the logic of an argument, the validity of a political or economic position, or the practicality of an idea. Brookfield (2012) and Fisher (2001) offer excellent explorations of critical thinking.
c. Decision-making is the process one goes through in order to make a decision. March (1994) provides a good primer on decision-making. Hoch and Kunreuther (2001) offer a thorough discussion.
d. Learning is the acquisition, understanding, and use of different types of knowledge effectively in differing circumstances. A learner must know and be proficient in using his or her primary and secondary learning styles; recognize when to use a particular learning style; be adept at sorting, storing, and managing information for easy retrieval; and competently practice self-management strategies to accomplish learning goals. An effective learner knows how he or she thinks and uses that information to efficiently and effectively learn.
e. Organizing is the process of bringing order out of disorder; it requires the intellectual skills of analysis and synthesis at a minimum. Consider a student preparing a class paper. The student picks a topic and then gathers information from different sources; next, his or her task is to sort through this information and form it into a coherent paper. Analysis is demonstrated when it is determined what information is needed; synthesis is used when that information is organized into a coherent paper. Evaluation may be required as well if the paper assesses another's opinion on perhaps a social issue or an argument favoring a business strategy.
f. Planning is the thinking skill that enables one to "map out" how to complete a task before starting. A student preparing a project must determine the supplies needed, lay out a schedule, anticipate and solve problems/roadblocks, manage time, and adjust the plan, as needed to accomplish the goal. The intellectual skills involved may include application, analysis, and/or synthesis.
g. Problem-solving involves problem analysis, generating solution alternatives (synthesis), and applying the selected alternative to resolve the presenting problem, and then assessing impact (evaluation). Hurson (2008) provides guidance on innovative problem solving. Zeitz (2007) presents a detailed treatment of problem solving.
h. Reasoning is use of deductive (moving from the specific to the general, such as concluding how all members of a group think, based on conversations with 3 or 4 group members) or inductive (going from the general to the specific, where the initial premise or assumption must be correct) reasoning to arrive at a conclusion. Royal (2010) provides an analysis of various reasoning skills. Holyoak and Morrison (2012) offer a substantial treatment of thinking and reasoning.

3. Three specific thinking skills are examined in detail (reasoning, critical thinking and decision-making) to show that their usage, combination of higher order intellectual skills and/or other thinking skills depends on the purpose of the brief or extended response test item, project, or work product used to assess learning target or benchmark mastery.
   a. Reasoning
      (1) Suppose a class assignment or end of course project requires learners to use "reasoning." What is meant by "reasoning" must be defined by the instructional designer, instructor, and/or examiner, who might be one person discharging each role; the point being is that "reasoning" must be consistently defined across course or module design, delivery, and assessment. Let's examine a framework developed by Quellmalz and Hoskyn (1997) as an example.
      (2) After reviewing the literature on frameworks for conceptualizing reasoning, Quellmalz and Hoskyn (1997) presented four reasoning skills: analysis, comparison, inference and interpretation, and evaluation.
         (a) Analysis is much the same as described by Bloom, et al. (1956). When a whole is divided into its component elements, relationships among and between those parts and their whole emerge. McMillan (2004, p. 172) points out that examinees, who are able to analyze, can "break down, differentiate, categorize, sort and subdivide."
         (b) Comparison entails the identification of differences and similarities. The learner compares, contrasts, or relates between and among explanations, data, arguments, assertions, or other objects of interest.
         (c) Inductive and deductive thinking gives rise to inference making (e.g., hypothesizing, generalizing, concluding, and predicting) and interpretation. We first make inferences and then interpret them.
         (d) Evaluation according to Quellmalz and Hoskyn (1997) is very similar to critical thinking. See Paul and Elder (2010) for an easy to digest, practical discussion.

   b. Critical Thinking
      (1) Ennis (1987, p. 10) defined critical thinking as "reasonable reflective thinking that is focused on deciding what to believe or do." Critical thinking is the ability to evaluate information, evidence, action, or belief in order to make a considered judgment as to its truth, value, and relevance. To assess critical thinking skills, interactive multiple choice exercises, extended response essays, and performance assessments are most suitable.

(2) An adaptation of Ennis' (1987) critical thinking approach is:
   (a) Clarify the problem, issue, or opportunity. Formulate an inquiry (e.g., proposition or question) within a relevant context. Ask questions or collect information which helps to clarify the problem, issue, or opportunity.
   (b) Collect more information. Assess the accuracy of facts and claims made by information sources. Distinguish between relevant and irrelevant information, arguments, or assertions. Detect bias in explanations, facts presented, arguments, or assertions made by information sources.
   (c) Apply inductive and deductive reasoning to the information collected. Identify logical inconsistencies and leaps in deductive and inductive reasoning from, within, between, and among the explanations, facts presented, arguments, or assertions made by information sources.
   (d) Analyze and synthesize the collected information. Search for implied or unstated assumptions; vague or irrational explanations, arguments or assertions; stereotypes; or name calling. Determine the types of critical relationships (e.g., coincidental, cause and effect, or spurious).
   (e) Make a judgment. Formulate alternative answers, solutions, or choices. Within the most suitable mix of costs, values, beliefs, laws, regulations, rules, and customs, consider each alternative and its anticipated consequence. Make a judgment, but be prepared to justify, explain, and argue for it. See Paul and Elder (2010) for an easy to digest, practical discussion.

c. Decision Making
   (1) A decision-making process may be diagramed as found in Figure 1.2. This is an example of procedural knowledge. Following the sequenced steps in a recipe is another example of procedural knowledge.

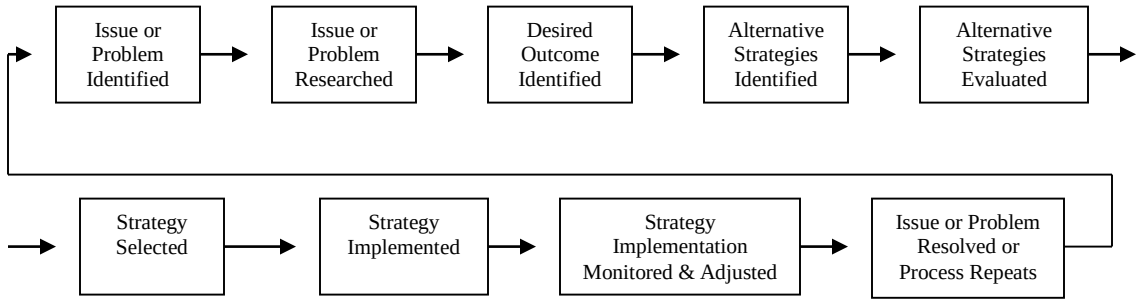

Figure Appendix 5.2.2 Decision-Making Process

(2) Decision-making Process
   (a) The first two stages require the decision-maker to identify the existence of an issue, problem, or opportunity and then research its causes, reasons for current existence, and impact.
   (b) Next, the decision-maker identifies his or her desired outcome.
   (c) Several strategies for attaining it are identified and evaluated as to its likelihood of success in producing the desired outcome.
   (d) Once alternative strategies are evaluated, one may be selected and implemented. If it is determined that no feasible corrective solution strategy exits, the decision-maker may stop the process.
   (e) Assuming a feasible corrective strategy is found, it is implemented, monitored and adjusted, as necessary.
   (f) After some predetermined time, cost, or other criteria, the issue or problem is declared resolved. If not resolved, the decision-making process repeats.

References

Anderson, L.W., & Krathwohl, D. (Eds.) (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Alexander, P. A. (1996). The past, present, and future of knowledge research: A re-examination of the role of knowledge in learning and instruction. *Educational Psychologist, 31*, 89-92.

Bloom. B. S., Engelhart, M. D., Frost, E. J., & Krathwohl, D. (1956). *Taxonomy of educational objectives. Book 1 cogitative domain*. New York, NY: Longman.

Brandt, R. S. (1986). On creativity and thinking skills: A conversation with David Perkins. *Educational Leadership, 43*(8), 12.

Brookfield, S. D. (2012). *Teaching for critical thinking*. San Francisco, CA: Jossey-Bass.

De Bono, E. (1999). *Six thinking hats*. New York, NY: Little, Brown & Company.

Depth of knowledge levels. (2005). Retrieved from http://dese.mo.gov/divimprove/sia/msip/DOK_Chart.pdf

Ennis, R. (1987). A taxonomy of critical thinking dispositions and abilities. In J. Barton & R. Sternberg (Eds.), *Teaching thinking skills* (pp. 9-26). New York, NY: W. H. Freeman and Company.

Farnham-Diggory, S. (1994). Paradigms of knowledge and instruction. *Review of Educational Research, 64*, 463-477.

Fisher, A. (2001). *Critical thinking: An introduction*. Cambridge, England: Cambridge University Press.

Gagne, R. M. (1985). *The conditions of learning and theory of instruction* (4th ed.). Chicago, IL: Holt, Rinehart, & Winston, Inc.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-46). New York, NY: Macmillan.

Hergenhahn, B. R. & Olson, M.H. (1997). *An introduction to theories of learning* (5th ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Hess, K. K. (2006). *Exploring cognitive demand in instruction and assessment* Retrieved from http://secondaryinstruction.muscogee.k12.ga.us/InstructionalSupport/DOK_Applying Webb_KH08%20Levels%20of%20Cognitive%20Demand%20p5.pdf

Hoch, S. J & Kunreuther, H. C. (2001). *Wharton making decisions*. New York, NY: John Wiley & Sons, Inc.

Holyoak, K. A. & Morrison, R. C. (2012). *The Oxford handbook of thinking and reasoning*. New York, NY: Oxford University Press.

Hurson, T. (2008) *An innovator's guide to productive problem-solving*. New York, NY: McGraw Hill.

Kagan, S. (2003). *Kagan structures for thinking.* Retrieved from http://www.kaganonline.com/free_articles/dr_spencer_kagan/ASK22.php

Kimble, G. A. (1961). *Hilgard and Marquis' conditioning and learning* (2nd ed.) Englewood Cliffs, NJ: Prentice Hall.

Kubiszyn, T. & Borich, G. (1996). *Educational testing and measurement*. New York, NY: Harper Collins College Publishers.

Levels of thinking in Bloom's taxonomy and Webb's depth of knowledge (n.d.). Retrieved from http://www.paffa.state.pa.us/PAAE/Curriculum%20Files/7.%20DOK%20Compared %20with%20Blooms%20Taxonomy.pdf

March, J. G. (1994). *A primer on decision-making*. New York, NY: The Free Press.

McMillan, J. H. (2004). *Classroom assessment: Principles and practice for effective instruction* (3rd ed.). Boston: Pearson.

Michalko, M. (1991). *Thinkertoys*. Berkeley, CA: Ten Speed Press.

Michalko, M. (2001). *Cracking creativity*. Berkeley, CA: Ten Speed Press.

Paris, S. G. & Cunningham, A. E. (1996). Children becoming students. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 117-146). New York, NY: Macmillan.

Paris, S. G. Lipson, M. Y. & Wixson, K. K. (1993). Becoming a strategic reader. *Contemporary Educational Psychology, 8*, 293-316.

Paul, R. & Elder, L. (2010). *The miniature guide to critical thinking concepts and tools*. Dillon Beach, CA: Foundation for Critical Thinking Press.

Quellmalz, E. S. & Hoskyn, J. (1997). Classroom assessment of reasoning strategies. In. G. D. Phye (Ed.), *Handbook of classroom assessment* (pp. 103-130). San Diego, CA: Academic Press.

Royal, B. (2010). *The little blue reasoning book*. Calgary, Alberta: Maven Publishing.

Sale, D. (n.d.). *Assessing specific types of thinking in problem-based learning activities.* Retrieved from http://www.tp.edu.sg/pbl_dennissale.pdf

Webb's Depth of Knowledge Guide. (2009). Retrieved from http://www.aps.edu/rda/documents/resources/Webbs_DOK_Guide.pdf

Woolfolk, A. (2001) *Educational psychology* (8th ed.). Boston, MA: Allyn & Bacon.

Zeitz, P. (2007). *The art and craft of problem solving* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

Appendix 5.2 Interpretative Exercise Example

The following essay was written by a student and is in rough-draft form. Read "Mt. Washington Climb." Then answer multiple-choice questions 15 through 18.

Mt. Washington Climb

1	I had gotten up around four-thirty that cool summer morning, anxious to get on the road.
2	My parents were still asleep, however, my mother had my lunch packed and waiting on the
3	kitchen table. I double-checked my gear, grabbed my boots, backpack, and lunch, and left
4	the house. Walking to my car, the first pale hint of red was entering the sky as the sun rose
5	higher and higher. I was right on schedule and greatly anticipating a good day.
6	It was a hot, dry summer morning when I reached the parking lot designated for hikers.
7	The cars in the parking lot were mostly from other states. I opened the trunk of my car and
8	grabbed my hiking boots. I had definitely chosen the right day for my first climb of Mt.
9	Washington. I tied my boots, put on my pack, made sure the car was locked up, and walked
10	over to the map displayed at the head of the trail. I studied the map for a few minutes and
11	then started my six-mile journey to the summit.
12	For the first two miles I walked slowly, enjoying the scenery. It was very beautiful out
13	there. The birds were out, the air was crisp and clean, and a soft breeze tickled my ears. After
14	an hour and a half had passed, I gradually picked up the pace.
15	I reached an intersection in the trail at about eleven o'clock. The sun was almost
16	directly overhead, and I judged the temperature as about ninety degrees. I had three miles
17	to go and I felt great. I drank a bottle of water before continuing. As I was about to.
18	get up, a huge deer walked right out in front of me; I never even heard it. It was by
19	far the most magnificent-looking animal I had ever seen. The deer's fur was light brown,
20	almost tan, and its antlers had strips of golden velvet that were torn and dirty. Just as
21	soon as the deer was there, he was gone, and I was back on my way to the summit
22	I walked cautiously among the trees for another hour. As I was walking, I noticed the
23	sky got brighter and brighter. Soon I broke through the tree line, and I could see the summit.
24	The sun glistened off the tower. Hundreds of people climbing toward the summit. I hesitated
25	for a moment, awed by the view, and then scrambled over the rocks toward the summit.
26	Beads of sweat ran down my face as I pushed toward the top. The summit was half a
27	mile away, yet it seemed like two feet. My legs burned but nothing could have stopped
28	me—not a rockslide, an earthquake, an avalanche—nothing. Determination filled my body
29	and gave me phenomenal energy. What seemed like two minutes was forty-five and before I
30	knew it I was on my knees at the summit. I had made it.

1. What is the correct punctuation edit for the sentence in lines 2 and 3 (My . . . table.)?
   A. My parents were still asleep; however, my mother had my lunch packed and waiting on the kitchen table.
   B. My parents were still asleep however; my mother had my lunch packed and waiting the kitchen table.
   C. My parents were still asleep, however, my mother had my lunch; packed and waiting on the kitchen table.
   D. My parents were still asleep, however, my mother had my lunch packed and waiting; on the kitchen table.

   ID:27480 Mt. Washington A

2. Which is the correct revision of the sentence in lines 4 and 5 (Walking . . . higher.)?
   A. The first pale hint of red was entering the sky as the sun rose higher and higher, walking to my car.
   B. As the sun rose higher and higher, walking to my car, the first pale hint of red was entering the sky.
   C. Walking to my car, I noticed the first pale hint of red entering the sky as the sun rose higher and higher.
   D. Walking to my car, the sun rose higher and higher as the first pale hint of red was entering the sky.
   ID:88949 Mt. Washington C

3. Which is a sentence fragment?
   A. I opened the trunk of my car and grabbed my hiking boots. (lines 7–8)
   B. It was very beautiful out there. (lines 12–13)
   C. Hundreds of people climbing toward the summit. (line 24)
   D. I had made it. (line 30)
   ID:27490 Mt. Washington C

4. Which sentence does not add to the development of the essay's main idea?
   A. The cars in the parking lot were mostly from other states. (line 7)
   B. I had definitely chosen the right day for my first climb of Mt. Washington. (lines 8–9)
   C. After an hour and a half had passed, I gradually picked up the pace. (lines 13–14)
   D. Beads of sweat ran down my face as I pushed toward the top. (line 26)
   ID:27487 Mt. Washington A

Source: New Hampshire Educational Assessment and Improvement Program, New Hampshire Department of Education. End-of-Grade 10 COMMON ITEMS, Released Items 2000-2001. Retrieved June 1, 2002 from: http://www.ed.state.nh.us/Assessment/2000-2001/2001Common10.pdf

Appendix 5.3 Task Specific, Analytical Scoring Rubric

Direct Performance Assessment Task Description

Student work teams will construct a direct performance task which meets the standards in the attached Performance Task Quality Assessment Index (PTQAI).

Each work team will develop a unit or mid-term performance task with a scoring rubric suitable for the 6th - 12th grade classroom. For this task, your work team will rely on the text, Internet research, team member experience, external consultants (e.g., senior teachers, etc.) and research and/or professional journals. The final work product is to be submitted electronically using Microsoft Word 2007 or a later edition.

The professor will give one "free read," using the Performance Task Quality Assessment Index (PTQAI) before "grading." The work product will be assessed on four dimensions or traits: quality of the task description, clarity and relevance of the performance criteria, scoring rubric functionality, and task authenticity, i.e., how realistic the intended task is for examinees. The rubric is analytical during the formative phase of the task and holistic at "grading," as a total score is awarded. Maximum points are 132. See the PTQAI.

The task will be divided into two parts: task description and rubric. Students will submit a clear set of directions (PTQAI items 1-3), appropriate task description (PTQAI items 4-10), relevant performance criteria (PTQAI items 11-15), a suitable scoring rubric (PTQAI items 16-23), and be authentic (PTQAI items 24-33). The performance criteria may be embedded within the scoring rubric.

The team must describe the classroom performance assessment context, using the attached Direct Performance Assessment Task Scenario.

Direct Performance Assessment Task Scenario

Before constructing the direct performance task, complete the scenario description to set the context within which it is set. Provide the following information about examinees and their school setting by either filling in the blank with the requested information or checking a blank as requested.

Ensure that intended learning target content or skills are fully described; do not refer the professor to a web site or other document containing this information. Otherwise, the task will be returned until the learning target information is supplied as required.

Examinee Characteristics

1) Ages: _____ (Average Age) 2) Grade Level: _____ (Specify)

3) Ethnic Mix (name ethnicity, percent of class):

_____

_____

4) Free/reduced Lunch: _____% of class 5) LEP: _____% of class

6) Exceptionalities present in class: _____

7) Classroom: Regular Ed: _____ Inclusive: _____ (Check one)

8) School Setting: Urban____ Suburban_____ Rural_____ (Check one)

9) School Ownership: Public: _____ Private: _____ Charter: _____ (Check one)

10) School Type: Elementary: ____Middle: ____ High: ___ (Check one)

11) School Grade Range: _____ 12) School Size: _____

13) Other Relevant Context Setting Information:

_____

_____

Learning Target(s): Clearly state the intended learning target content and/or skills which are to be assessed in your extended performance task.

Classroom Assessment Plan: Describe in a few paragraphs your general classroom assessment plan. Next, describe how your team's extended performance task would fit into that assessment plan and how you would use the information to improve classroom teaching and learning.

Performance Task Quality Assessment Index (PTQAI)

The PTQAI is the scoring rubric used to assess the product to be produced in response to the task description. The PTQAI is composed of 33 evaluative criteria (called standards) distributed across these four subtests or dimensions: Task Description, Performance Criteria, Scoring System, and Authenticity.

Possible scores are "Meets Standard," 4 points: "Mostly Meets Standard," 3 points; "Marginally Meets Standard," 2 points; "Does Not Meet Standards," 1 point; and "Missing or Not in Evidence," 0 points. Performance level quality definitions are:

"Meets Standard" means the response fully and completely satisfies the standard.

"Mostly Meets Standard" means the response nearly fully and completely satisfies the standard.

"Marginally Meets Standard" means the response partly satisfies the standard.

"Does Not Meet Standard" means the response does not fully and completely satisfy the standard.

"Missing or Not in Evidence" means the response presents no evidence of meeting the standard.

In order to compute a total score for summative evaluation purposes, the following criteria are used:

Exceptional corresponds to an A (95-100%). Performance is outstanding; significantly above the usual expectations.

Proficient corresponds to a grade of B to A- (83-94%). Performance is at the level of expectation.

Basic corresponds to a C to B- (75-82%). Performance is acceptable but improvements are needed to meet expectations well.

Novice corresponds to an F (< 74%). Performance is weak; the skills or standards are not sufficiently demonstrated at this time.

To compute a summative performance level, total earned points are divided by points possible (132).

Performance Task Quality Assessment Index (PTQAI)	
Each direct performance task should meet the 33 standards presented in this index. Four dimensions of performance assessment are examined: the task description, performance criteria, the scoring system, and authenticity. For each standard, one of five performance levels or scores is possible: 0= Missing, 1= Does not meet standard, 2=Marginally standard, 3= Mostly meets standard or 4= Meets standard.	
Task Description	**Score**
1. Directions are explicitly stated.	
2. Directions are likely to be understood by examines.	
3. Directions are likely to produce intended examinee behavior.	
4. Critical learning target content/skills are integrated into the task.	
5. The task allows for multiple plausible solutions.	
6. Expected examinee activities (e.g., individual or group effort, internet research, interviews, etc.) required to complete the task is explicitly stated.	
7. Resources needed to complete the task are explicitly stated.	
8 Expected performance or work products are so described to enable examinee understanding.	
9. The teacher's role in relation to the task is fully described.	
10. Scoring procedures are so described to enable examinee understanding.	
Performance Criteria	**Score**
11. Performance criteria are explicitly stated (i.e., unambiguous).	
12. Performance criteria are likely to be understood by examinees.	
13. Performance criteria focus on important task dimensions.	
14. Performance criteria are logically related to the task.	
15. Performance criteria are directly observable, little or no inference making.	
Scoring System (Rubric)	**Score**
16. The type of rating scale (holistic or analytical) is appropriate.	
17. The scoring procedure is feasible, i.e., workable.	
18. The scoring procedure minimizes scoring error.	
19. The scoring procedure is likely to be understood by examinees.	
20. Rating scale increments (e.g., numerical or qualitative) are suitable.	
21. Performance level descriptions are logically related to the task.	
22. Performance level descriptions are suitable given the task.	
23. Performance level descriptions are likely to be understood by examinees.	
Authenticity	**Score**
24. The task is feasible, but reasonably challenging for examinees.	
25. The task requires examinees to employ suitable intellectual skills (e.g., analysis, synthesis).	
26. The task requires examinees to use apt thinking skills (e.g., creativity, problem solving, etc.).	
27. Examinees must apply learning target content and/or skills.	
28. The task replicates or simulates an academic, personal, civic, or vocational event, experience, etc. examinees are likely to encounter.	
29. Examinees are required to efficiently and effectively apply learning target content and/or skills to a suitably complex task.	
30. Examinees must rehearse or practice required learning target content and/or skills to successfully complete the task.	
31. Examinees must consult resources (e.g., texts, internet, libraries, people, etc.) to complete the task.	
32. Examinees are provided feedback before "graded" final performance or work product submission.	
33. Examinees must complete the task under suitable constraints (e.g., time, prior knowledge, resources, etc.).	
Total Score (Maximum = 132)	
Comments:	

Appendix 5.4 Group Contribution Rating Scale

Read each statement carefully. Next, circle either 2 for "Yes" or 1 for "No" if the indicator (behavior) was demonstrated by the group member or not. The "?" indicates that the indicator was not observed.

Standard (Evaluative Criterion)	**No**	**Yes**	**?**
1. The group member's participation was focused on the task at hand.	1	2	0
2. The group member usually exhibited a respectful demeanor.	1	2	0
3. The group member contributed an acceptable quantity of data, e.g., research articles, URLs, books, etc., given the team's task.	1	2	0
4. The quality of the group member's data (e.g., research articles, URLs, books, etc.) contribution was high, given the task.	1	2	0
5. The group member's contribution of data (e.g., research articles, URLs, books, etc.) was relevant to the team's task.	1	2	0
6. The group member acceptably met the team's deadlines.	1	2	0
7. When required, the member exhibited appropriate mediating skills.	1	2	0
8. The member followed team directions in an acceptable manner.	1	2	0
9. The group member exhibited appropriate listening skills which assisted the team in accomplishing its task.	1	2	0
10. The team member was sufficiently flexible so as to enable the work group to complete the task at hand.	1	2	0
11. The team member demonstrated writing skills, which helped the work group meet its objective.	1	2	0
12. By providing constructive feedback to team mates, the member contributed towards accomplishing the team's task.	1	2	0

Total Earned Points:_____

Interpretation: The higher the number of points, the higher the perceived contribution level.

Appendix 5.5 Skill Focused Scoring Rubric
Article Review Scoring Rubric

Names _____ Date Completed_____ Total Score_____ of 100

Students will work in groups to review and critique a short article from a peer-reviewed **electronic** journal which reports on a single quantitative, qualitative, or mixed-method study. Do not select a literature review or a meta-analysis article.

The group will critique the work based on evidence of reliability, validity, design suitability, and practical usefulness of the information. When asked to discuss the suitability and sufficiency of any supporting data, briefly summarize the data and then critique it using prevailing best practices, professional standards, your text, and other research. Cite references in an APA style reference list. The task is a three page maximum, excluding title and reference pages. Left – margin headings: Intent, Type, Reliability/Authority, Validity/Verisimilitude, and Conclusion. **The article's URL must be provided.**

Rating:

Exceptional corresponds to an A (95-100%). Performance is outstanding; significantly above the usual expectations.
Proficient corresponds to a grade of B to A- (83-94%). Skills and standards are at the level of expectation.
Basic corresponds to a C to B- (75-82%). Skills and standards are acceptable but improvements are needed to meet expectations well.
Novice corresponds to an F (< 74%). Performance is weak; the skills or standards are not sufficiently demonstrated at this time.
0 This criterion is missing or not in evidence.

Criteria	Ratings				
	0	**Novice**	**Basic**	**Proficient**	**Exceptional**
Intent. The intent of the research is summarized succinctly and thoroughly in a style appropriate to the research design. State the purpose for which the research was conducted.		1.0 - 7.4	7.5 – 8.2	8.25 – 9.4	9.5 - 10
Type. State whether the study was primarily or exclusively quantitative, qualitative, or mixed method. Briefly describe the study's methodology to prove your designation.		1.0 - 7.4	7.5 – 8.2	8.25 – 9.4	9.5 - 10
Reliability/Authority. For an article which reports on a quantitative study, describe the data collection device's internal consistency, stability, equivalence, or inter-rater reliability; give coefficients if available. For an article reporting on a qualitative study, describe the author's qualifications and experience, the sponsoring organization, number of times the article has been cited in other research, cite other researcher's opinions of the article, how consistent the reported findings are with other studies, etc.) Discuss the suitability and sufficiency of any supporting data.		1.0 – 18.4	18.5 – 19.9	20 - 23.4	23.5 - 25
Validity/Verisimilitude**.** For an article which reports on a quantitative study, describe the data collection device's content, criterion-related, or construct validity; also comment on the control or prevention of internal research design threats to internal design validity. For an article reporting on a qualitative study, show the article's verisimilitude (i.e., appearance of truth). Do this by commenting on the logical analysis used by the authors; describe how consistent their findings and recommendations are with other researchers; comment on the consistency of the study's design and execution with other research on the topic, etc. When assessing logical analysis, consider logic leaps, the internal consistency of arguments, deductive and inductive reasoning, etc. Discuss the suitability and sufficiency of any supporting data.		1.0 – 18.4	18.5 – 19.9	20 - 23.4	23.5 - 25
Conclusion. Provide an overall assessment of research reported in the article. Did the study meet prevailing best practices for its research design and data collection strategies? Why or why not? Describe the practical application of findings to professional practice.		1.0 – 11.4	11.5 – 12.4	12.5 – 14.4	14.5 - 15
Writing and grammar skills are appropriate to the graduate level (including APA citations and references).		1 – 11.1	11.2-12.3	12.4-14.1	14.2-15
					Total Earned Points: _____

Appendix 5.6 Presentation Rating Rubric

The presentation scoring rubric is presented below and is composed of three subtests or sections: Organization, Presence, and Technology. Odd number ratings reflect the "mid-point" between two even numbered scores.

A. Execution: Score ____ (out of 48)

1. **Introduction: Score ____ (out of 6)**
   - 0 – Poor to nonexistent introduction, no "attention grabber"
   - 2 – Vague objectives or simply reads the problem statement
   - 4 -- Good "attention-grabber", weak objective foundation
   - 4 -- Weak "attention-grabber", clear objective foundation
   - 6 – Good "attention-grabber", lays clear foundation of objectives

2. **Content: Score ____ (out of 8)**
   - 0 – Wrong data, wrong problem, "great" leaps in logic, poor question anticipation
   - 4 – Some understanding of presentation content demonstrated, some incorrect terminology, a few small leaps in logic, fair question anticipation
   - 6 -- Proficient understanding of presentation content demonstrated, little incorrect terminology, no leaps in logic, good question anticipation
   - 8-- Exemplary understanding of presentation content demonstrated, no incorrect terminology, no leaps in logic, excellent question anticipation

3. **Questioning: Score ____ (out of 6)**
   - 0 – Failed to answer questions, engaged in no discussion
   - 2 -- Answered questions awkwardly, partial explanations, ineffective discussion
   - 4 -- Answered questions somewhat effectively, with accurate explanations, poorly lead discussion(s)
   - 6 -- Answered questions effectively with accurate explanations & effectively lead discussion(s)

4. **Communication Strategy: Score ____ (out of 6)**
   - 0 -- Inappropriate strategy, ineffectually executed
   - 2 -- Appropriate strategy, but ineffectually executed
   - 4 -- Appropriate strategy, proficiently executed
   - 6 -- Appropriate strategy, expertly executed

5. **Use of Notes: Score ____ (out of 6)**
   - 0 -- Relies heavily on notes or prepared text, lost place several times
   - 2 -- Relies moderately on notes or prepared text, lost place some times
   - 4 -- Relies little on notes or prepared text, lost place once or twice
   - 6 -- Doesn't rely on notes or prepared text, lost place only once

6. **Conclusion and Wrap-Up: Score ____ (out of 8)**
   - 0 – Poor to nonexistent conclusion
   - 2 – Simple re-hash of main point(s), and "not tied together"
   - 4 – Main point(s) clarified, implications weakly presented, & "tied together"
   - 6 --Main point(s) clarified, implications reasonably well presented & "tied together"
   - 8 --Main point(s) clarified, implications well presented & "tied together"

7. **Professional Impression: Score ____ (out of 8)**
   - 0 -- "Slip-shod" appearance
   - 2-- Amateurish in appearance
   - 4 -- Less amateurish in appearance
   - 6 -- Proficient in appearance
   - 8 -- Expert in appearance

B. Presence: Score ____ (out of 24)

1. **Eye Contact: Score ____ (out of 4)**
   - 0 – Makes little eye contact,
   - 2 – Makes moderate eye contact, focuses on one group or side of the room
   - 4 – Makes and holds eye contact with people all over the room,

2. **Use of Hands and Body Movement: Score ____ (out of 4)**
   0 – Distracts or annoys audience or gives perception of being nervous
   2 – Somewhat comfortable but movements interrupt flow of presentation
   4 –Completely comfortable, appropriate hand-gestures, and non-awkward movements

3. **Voice & Inflection: Score ____ (out of 6)**
   0 – Too hard to hear, sounds disinterested, did not project voice
   2 – Speaks in a monotone, sounds disinterested, or many "ums" and "likes", inconsistent voice projection
   4 -- Varies voice and inflection appropriately, conveys some enthusiasm, fairly consistent voice projection
   6 – Varies voice and inflection expertly, conveys enthusiasm, appropriate voice projection

4. **Articulation & Pace: Score ____ (out of 4)**
   0 -- Poorly articulated words and/or sentences, very distracting speaking pace, many awkward pauses
   2 -- Mispronounces some words and/or "mangles" sentences, inconsistent speaking pace, few awkward pauses
   4 -- Articulates words and sentences clearly, speaking pace appropriate, no awkward pauses

5. **Professional Appearance: Score ____ (out of 6)**
   0 -- Very inappropriately attired for presentation subject, audience, and/or environment
   2 -- Somewhat inappropriately attired for presentation subject, audience, and/or environment
   4 -- Appropriately attired for presentation subject, audience, and/or environment
   6 -- Very appropriately attired for presentation subject, audience, and/or environment

C. Technology: Score ____ (out of 28)

1. **Slides, Graphics, Figures, etc. Layout: Score ____ (out of 4)**
   0 -- Visual aids are poorly designed, cluttered, & many have missing labels
   2 -- Visual aids are sometimes difficult to read; some slides are cluttered, and labeling was inconsistent.
   4 -- Visual aids easy to read, uncluttered, and fully labeled

2. **Slides, Graphics, Figures Color: Score ____ (out of 4)**
   0 – Visual aid colors distracting and confusing for presentation
   2 -- Visual aid color appropriateness inconsistent for presentation
   4 -- Visual aid colors appropriate for presentation

3. **Slide, etc. /Presenter Alignment: Score ____ (out of 4)**
   0—Visual aids and presenter were frequently out-of-alignment
   2—Visual aids and presenter were occasionally out-of-alignment
   4—Visual aids and presenter were rarely out-of-alignment

4. **Slide, etc. Reading: Score ____ (out of 4)**
   0—Presenter frequently read slides to audience
   2—Presenter occasionally read slides to audience
   4—Presenter rarely read slides to audience

5. **Slide Presentation Support: Score: _____ (out of 4)**
   0—Visual aids ineffective & hard to follow
   2— Visual aids moderately effective & easy to follow
   4— Visual aids effective & easy to follow

6. **General Usage: Score ____ (out of 4)**
   0 -- Visual aids lacking or poorly utilized, very distracting
   2 -- Visual aid effectiveness inconsistent and distracting, at points, during presentation
   4 -- Visual aids utilized effectively and not distracting

7. **Usage Impact: Score ____ (out of 4)** **Total Score: _____**
   0 -- Learning not enhanced
   2 -- Learning enhancement inconsistent
   4 -- Learning enhanced

Appendix 5.6: Preparing Examinees for a Testing Session

A. Preparing the Testing Environment

1. Each time an examination is administered, the testing environment should be similar to that of a nationally standardized test, unless a performance assessment is administered; modify these guidelines as appropriate.
   a. The room should be clean, uncluttered, and ordered.
   b. The temperature should be comfortable, adequately lighted, and well ventilated.
   c. Test materials should be ready and in sufficient number.
   d. Distractions such as noise or unnecessary movement should be avoided.
   e. Discipline should be maintained.
   f. Answer individual questions about items carefully. This can be substantially reduced if language and item construction are developmentally appropriate.

2. Examinees are going to be anxious and some will be frustrated; reassure as best as is possible. Breathing exercises may be helpful as will the progressive relaxation of muscles. If an examinee's text anxiety appears to be chronically acute, refer the examinee for learning disability evaluation.

3. Special Learning Assessment Applications
   a. Open book tests are helpful if the testing emphasis is on application, not memorization.
   b. Unannounced tests are not recommended as examinees tend to under-perform due to anxiety and inadequate preparation time. Study time + test preparation = high scores.
   c. For summative evaluation, selection, and placement, single test administrations are sufficient. For diagnostic and formative evaluation, frequent testing is recommended. Frequent testing motivates examinees and improves scoring.

B. Ethical Test Preparation

1. Test-preparation training tends to produce higher test scores. To improve test-taking skills,
   a. Examinees need to understand the mechanics of test-taking, such as the need to carefully follow instructions, checking their work, and so forth.
   b. Examinees should use appropriate test-taking strategies, including ways in which test items should be addressed and how to make educated guesses.
   c. Examinees need to practice test-taking skills.

2. Successful examinees tend to understand the purpose for testing, know how the results will be used, comprehend the test's importance and its relevance to teaching and learning, expect to score well, and have confidence in their abilities.

3. Test preparation is effective. A critical question is, "When does test preparation become teaching to the test?"
   a. Mehrens and Kaminski (1989) suggest that providing general instruction on content and performance objectives without specific reference to those tested by an intended achievement (often, a standardized) test, and the teaching of test taking skills, is ethical.
   b. They argue that:
      (1) Teaching content and performance standards which are common to many achievement tests (standardized or locally developed);
      (2) Teaching content and performance standards which specifically match those on the test to be administered; and
      (3) Teaching to specifically matched content and performance standards and where practice follows the same item formats as found on the test to be administered is a matter of professional opinion.
   c. However, provision of practice or instruction on a parallel form of the test to be administered or on the actual test itself is unethical (Mehrens & Kaminski, 1989).
   d. Clearly, the "cut" lies between b(1), b(2), or b(3). Two guiding principles should be employed.
      (1) The justification for teaching test taking skills is that such instruction reduces the presence of random error in an examinee's test score so that a more accurate measure of learning is taken. It seems that the teaching of specific content known to be included on a test to be administered and which is not part of the student's regular instructional program is teaching to the test and thus artificially inflating the examinee's test score. There should be very tight alignment between any standardized test and the regular instructional program or the standardized test should not be used to measure achievement.
      (2) The second guideline is the type of test score inference to be drawn. Mehrens & Kaminski (1989) write, "the only reasonable, direct inference you can make from a test score is the degree to which a student knows the content that the test samples. Any inference about why the student knows that content...is clearly a weaker inference."
         (a) Teaching to the test involves teaching specific content which weakens the direct inference about what the examinee knows and can do. Testing is done to generalize to a fairly broad domain of knowledge and skills, not the specific content and item format presented on a specific test.
         (b) When one seeks to infer from a test score why an examinee knows what he or she knows, an indirect inference is being made. Indirect inferences are dangerous and often lead to interpretation errors.
   e. Applying these two guidelines leads one to conclude that
      (1) Providing general instruction on content and performance objectives without reference to those tested by an intended achievement (often, a standardized) test;

(2) Teaching test taking skills; and/or
(3) Teaching content and performance standards which are common to many achievement tests (standardized or locally developed) are ethical.

C. Test Preparation Principles Performed by Teachers for Students

1. Specific Classroom Recommendations are:
   a. If practice tests are used in the examinee's classroom, make it a learning experience. Explain why the correct answer is correct and why the incorrect answers are incorrect. Use brainstorming and other strategies that promote a diversity of responses. Examinees should develop questions for class discussions and practice tests. Have the class (as in small groups) identify content from the text, notes, supplemental learning materials etc. which points towards the correct answer.
   b. Incorporate all intellectual skills into daily activities, assignments, class discussion, homework, and tests. Teach examinees the different intellectual skills; help them learn to classify course or class activities and questions by intellectual skill so that the various categories of thinking (recall, analysis, comparison, inference, and evaluation) are learned and practiced.
   c. Encourage examinees to explain their thinking, i.e., how they arrived at their answer, conclusion, or opinion. Practice discerning fact from opinion and the relevant from the irrelevant. Practice looking for relationships among ideas by identifying common threads. Have examinees solve verbal analogies, logic puzzles, and other classification problems.
   d. Apply learned information to new and different situations or issues. Encourage application of information by asking examinees to relate what has been learned to their own experiences.
   e. Ask open-ended questions which ensure that examinees do not assume that there is one correct answer.
   f. Assign time limits to classroom work and structure assignments, quizzes, or tests in formats similar to those found on standardized tests.

2. Improving Examinee Motivation Recommendations
   a. Expect good results and model a positive attitude.
   b. Use appropriate motivational activities and provide appropriate incentives.
   c. Discuss the test's purpose and relevance and how scores are to be used.
   d. Discuss with examinees the testing dates, times, content, time available to complete the test(s), test item format, and length of reading passages if any.
   e. Know and use correct test administration procedures while ensuring a quiet, orderly testing environment.
   f. Ensure that all skills to be tested have been taught and practiced to proficiency. Ensure that examinees know they have adequately prepared for the testing experience. Test-preparation is not a substitute for thorough, diligent preparation.

D. Test Preparation Principles Performed by Students

1. List the major topics for the entire course (if a comprehensive midterm or final) or for chapter(s) which contribute test content. Review all of relevant source material: textbook(s), notes, handouts, outside readings, returned quizzes, etc.

2. For each topic, in an outline, summarize key or critical information. It will also be helpful to construct a map of the major concepts so that you will note connections between key concepts.

3. Plan study time. Study material with which you have the most difficulty. Move on to more familiar or easier material after mastering the more difficult content. However, balance time allocations so that there is sufficient time to review all critical or key material before the test.

4. Study material in a sequenced fashion over time. Allocate three or four hours each day up to the night before the test. Spaced review is more effective because you repeatedly review content which increases retention and builds content connections. "Don't cram." Develop a study schedule and stick to it. Rereading difficult content will help a student learn. Tutoring can be a valuable study asset.

5. Frame questions which are likely to be on the test in the item format likely to be found. If an instructor has almost always used multiple choice items on prior tests or quizzes, it is likely that he or she will continue to use that item format. Anticipate questions an examiner might ask. For example:
   a. For introductory courses, there will likely to be many terms to memorize. These terms form the disciplinary vocabulary. In addition to being able to define such terms, the examinee might need to apply or identify the terms.
   b. It is likely the examinee will need to apply, compare and contrast terms concepts or ideas, apply formulae, solve problems, construct procedures, evaluate a position or opinion, etc. Most likely, the examinee will need to show analytical and synthesis skills.

6. Form Study Groups
   a. Study groups tend to be effective in preparing examinees provided the group is properly managed. Time spent studying does not necessarily translate into leaning, and hence improved test scores.
   b. Study group benefits include shared resources, multiple perspectives, mutual support and encouragement, and a sharing of class learning aides (e.g., notes, chapter outlines, summaries, etc.).
   c. To form a study group, seek out dedicated students (i.e., those who ask questions in class and take notes), meet a few times to ascertain whether or not the group will work well together, and keep the group limited to five or six members.

d. Once started, meet regularly and focus on one subject or project at a time, have an agenda for each study session, ensure that logistics remain "worked out" and fair, follow an agreed upon meeting format, include a time-limited open discussion of the topic, and brainstorm possible test questions.

7. Other Recommendations
   a. For machine-graded multiple-choice tests, ensure the selected answer corresponds to the question the examinee intends to answer.
   b. If using an answer sheet and test booklet, check the test booklet against the answer sheet whenever starting a new page, column, or section.
   c. Read test directions very carefully.
   d. If efficient use of time is critical, quickly review the test and organize a mental test completion schedule. Check to ensure that when one-quarter of the available time is used, the examinee is one-quarter through the test.
   e. Don't waste time reflecting on difficult-to-answer questions. Guess if there is no correction for guessing; but it is better to mark the item and return to it later if time allows, as other test items might cue you to the correct answer.
   f. Don't read more complexity into test items than is presented. Simple test items almost always require simple answers.
   g. Ask the examiner to clarify an item if needed, unless explicitly forbidden.
   h. If the test is completed and time remains, review answers, especially those which were guessed at or not known.
   i. Changing answers may produce higher test scores, if the examinee is reasonably sure that the revised answer is correct.
   j. Write down formula equations, critical facts, etc. in the margin of the test before answering items.

E. Strategies for Taking Multiple Choice Tests

1. Read the question. Advise examinees to think of the answer first as he or she reads the stem. By thinking of the answer first, he or she is less likely to be fooled by an incorrect answer. However, read all answer options before selecting an answer.
2. Do not spend too much time on any one question, as the item may actually have two correct answers, instead of one or no correct answer.
   a. In these cases select an answer at random, provided there is no penalty for guessing and wrong answers are not counted.
   b. Advise the examinee to circle the question number so he or she can go later if there is time. Go on to the next item.
3. If an examinee doesn't know the answer, then he or she should mark out options which are known to be incorrect. This increases the chances of a correct guess. Many multiple choice items, have only two plausible answer options, regardless of the number presented.
4. Do not keep changing answers. If the item seems to have two correct answers, select the best option and move along.

   a. Based on testing research, the first answer is probably correct. An examinee is most likely to change a correct answer to an incorrect one.
   b. Only change an answer if absolutely sure a mistake was made.
5. After finishing the test, go back to circled items.
   a. Don't leave a testing session early, unless absolutely sure each item is correctly answered. Invest what time is available, to answer unanswered items.
   b. If an item is still not able to be answered, guess. You have a 25% chance of selecting the correct answer, if four-options are presented. The chances of a correct guess are higher if other answer options can be eliminated.
6. If the test is not properly constructed, the following "tricks" might raise scores:
   a. Where the examinee must complete a sentence, select the one option that fits better grammatically.
   b. Answer options which repeat key words from the stem are likely correct.
   c. Longer answer options tend to be correct more often than incorrect.
   d. If there is only one correct answer to an item, that answer is likely to be different. Thus if two or three answer options mean the same, they must be incorrect.
   e. If guessing and a typing error is noted in an answer option, select another option.
7. If two answer options have the same meaning, neither are likely to be correct.
8. If two answer options have opposite meanings, one is usually is correct.
9. The one answer which is more general than the others is usually the right answer.
10. For an occasional test item whose answer options end with "all the above"; select "all of the above". If one answer option is incorrect, then "all of the above" cannot be a correct option.
11. The answer option which is more inclusive, (i.e., contains information which is also presented in other answer options), is likely to be the correct.
12. When you have studied and don't know the answer, select "C" if there is no guessing penalty.
13. If you do not lose points for incorrect answers, consider these guidelines for making an educated guess:
   a. If two answers are similar, save for a couple of words, select one of them.
   b. If a sentence completion stem, eliminate any possible answer which would not form a grammatically correct sentence.
   c. If numerical answer options cover a wide range, choose a number in the middle.
14. Answer options containing always, never, necessarily, only, must, completely, totally, etc., tend to be incorrect.
15. Answer options which present carefully crafted statements incorporating such qualifiers as often, sometimes, perhaps, may and generally, tend to be correct.

F. Strategies for Answering Other Test Item Formats

1. True-False Tests
   a. If any part of a statement is false, the answer is false.
   b. Items containing absolute qualifiers, e.g., always or never, often are false.
2. Open Book Tests
   a. Write down any formulas you will need on a separate sheet.
   b. Place tabs on critical book pages.
   c. If using notes, number each page and make a table of contents.
   d. Prepare thoroughly; these types of examinations are often very difficult.
3. Short Answer/Fill-in-the-Blank
   a. These test items, ask examinees to provide definitions (e.g., a few words) or short descriptions in a sentence or two.
   b. Use flashcards with important terms and phrases, missing or highlighted when studying. Key words and facts will be familiar and easy to recall.
4. Essay Tests
   a. Decide precisely what the question is asking. If a question asks you to contrast, do not interpret.
   b. If an examinee doesn't know an answer or if testing time is short, usually it's a good idea to answer those items that the examinee knows the answers to first. Then sort based on a combination of "best guess" and available points. Attempt to answer those items worth the most points and for which the examinee has the greatest amount of knowledge.
   c. Verbs used in essays include: analyze, compare, contrast, criticize, define, describe, discuss, evaluate, explain, interpret, list, outline, prove, summarize. Look up any unfamiliar words in a dictionary.
   d. Before writing, make a brief, but quick outline.
      (1) Thoughts will be more organized and the examinee is less likely to omit key facts and/or thoughts.
      (2) The examinee will write faster and may earn some points with the outline, if he or she runs out of time.
      (3) Points are often lost as the examiner has little understanding of an examinee's response due to its poor organization. Use headings or numbers to guide the reader.
   e. Leave plenty of space between answers. The extra space is needed to add information if time is available.
   f. When you write, get to the point. Start off by including part of the question in your answer to help focus your response.
      (1) Build upon your answer with supporting ideas and facts.
      (2) Review your answers for proper grammar, clarity and legibility.
   g. Don't pad answers; this tends to irritate examiners. Be clear, concise, and to the point. Use technical language appropriately.

References

Mehrens, W. A. & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8,* (1), 14-22.