Part I The Bases for Assessment in the Classroom

A. Classroom Decision Making and Using Assessment (cf. 1,3, & 7)

1. Basic Concepts

a. Assessment: A broad and comprehensive term referring to a process for obtaining information used for making decisions about students, curricula, programs, and educational policy. Assessment may be either **formative** or **summative** (see d. Evaluation below).

b. Test: An instrument or systematic procedure for observing and describing one or more characteristics of a student using either a numerical scale or a classification scheme. More narrow than assessment.

c. Measurement: a procedure for assigning numbers (usually called scores) to a specified attribute or characteristic of a person in such a way that the numbers describe the degree to which the person possesses the attribute.

d. Evaluation: the process of making a value judgement about the worth of a studentÂ product or performance.
   1) Formative: judgement about the quality or worth made during the design or development of instructional materials, instructional procedures, curricula, or educational programs.
   2) Summative: judgement about the quality or worth of already-completed instructional materials, instructional procedures, curricula, or educational programs.

2. Assessment and Types of Educational Decisions about Students

a. Instructional Management Decisions: includes decisions that regard the instructional diagnosis and remediation of students, providing student feedback and teacher feedback, modeling learning objectives, motivating students, assigning grades to students.

b. Selection Decisions: includes decisions that regard which persons who are acceptable or not acceptable for admission into a class or program. It is important to show that the candidateÂ results on the assessments are related to success in the program or job for which the institution is selecting persons.

2. Assessment and Types of Educational Decisions about Students (Cont'd)

c. Placement Decisions: includes decisions that regard assigning people to different levels of the same general type of instruction, education, work; no one is rejected, but all remain within the institution to be assigned to some level. Again, unlike Selection Decisions, rejection is not possible. Most decisions in schools are placement decisions; students are often said to be , screened„.

d. Classification Decisions: includes decisions that regard assigning persons to several different, but unordered categories, jobs, or programs. Classifying students with disabilities as hearing-impaired or visually-impaired does not imply some order.

e. Counseling and Guidance Decisions: includes decisions that regard assisting students in exploring and choosing careers and in directing them to prepare for the careers they select. A series of assessments are usually given including an interview, an interest inventory, aptitude tests, a personality questionnaire, and an achievement battery (i.e., a group of related tests).

f. Credentialing and Certification Decisions: includes decisions that regard whether students have attained certain standards of learning (minimum standards, high standards). Credentialing and Certification may be voluntary or mandatory (as with state legislation).

3. Norm referenced and Criterion referenced Interpretations

a. Norm referenced interpretations: describe assessed performance in terms of a personÂ position relative to some standard reference group that has been previously administered the assessment. The reference group ideally should represent the general population of persons to which the person under review belongs. This group is usually called the **norm group**. A learning disabled child should be compared to a population of learning disabled children of approximately the same age, with the same disability when a norm referenced interpretation is to be made about the childÂ performance.

b. Criterion referenced interpretations: describe assessed performance in terms of the kinds of tasks a person with a given score can do.

c. General comment: Both kinds of interpretations are important to understand how well a student is learning.

4. Additional ways to describe assessments

a. Items: another name for questions, exercises, and tasks appearing on an assessment procedure. Items may be Response-choice, Completion, Short answer, or Constructed response

b. Objective Scoring versus Subjective Scoring: Objectivity is a matter of degree, with true-false and multiple-choice exams tending to be objectively scored and essay, portfolio, and performance assessments tending to be subjectively scored. Unlike Objective assessments, Subjective scored assessments notably have a history of being scored differently by different people or even differently by the same person over time.

c. Verbal versus Performance assessments: Verbal assessments are based upon the verbal responses of students. Verbal responses, oral or in writing, are, in the end, behaviors. So it is reasonable to speak of verbal behavior in contrast to a performance, in which students accomplish some task (assembling puzzles, building a tower of blocks, completing a chemistry experiment, or even throwing a football).

d. Standardization: the degree to which the observational procedures, administration procedures, equipment and materials, and scoring rules have been fixed so that, insofar as possible, the same procedure occurs at different times and places.

e. Power Assessments: Assessments in which time limits are generous because the focus is on assessing the amount of knowledge, comprehension, or understanding a student possesses.

f. Speeded Assessments: Assessments in which time limits are restricted because performance speed in answering questions or accomplishing a task is a key focus.

g. Interest assessments: assess preference for particular activities

h. Value assessments: assess preference for ,life goals„ and ,ways of life„.

i. Attitude assessments: assess feelings about particular social objects-physical objects, types of people, particular persons, social institutions, government polices, and others.

B. Describing the Goals and Learning Targets of Instruction (cf. 1,3, & 7)

1. Basic Concepts

a. Student Learning Outcome Objectives (targets): Specifies what you would like the students to be able to do, value, or feel at the completion of an instructional segment. (Variably called **learning objectives**, outcome objectives, or learning targets)

b. Instruction: The process used to provide students with the conditions that help them achieve the learning objectives. Instruction involves three interrelated activities: (1) deciding what the students are to learn, (2) instructing the students, and (3) evaluating the learning that takes place. Most teachers find that these three activities do not follow a clear linear order and that the process is truly cyclical. Generally, the progression from one activity to the next follows this sequence, though often it is pedagogically important to revise each step upon each cycle of instruction, when the material is taught anew.

c. Learning Objectives may be classified into three groups: Cognitive, affective (e.g., emotional or value-oriented) and psychomotor (e.g., performance oriented).

d. Principles to consult when crafting Learning Objectives

1. Student-Centered: State in terms of intended student outcomes. State the objectives in terms of how the STUDENTS participating in your class will learn, NOT so much what you accomplished or taught.

2. Content-Centered Make sure you indicate the content to which the objective applies by explicitly referring to the specific materials to be learned.

3. Content-Centered: Identify objectives that completely represent each of your identified goals.

4. Content-Centered: Enumerate your objectives, one concept for one objective. Avoid complex sentences. A sentence with more than one idea can be broken down into more than one objective. Be specific and clear.

5. Content-Centered: Make sure people outside of your discipline can read the objectives and understand them.

d. <u>Principles to consult when crafting Learning Objectives</u> (Cont'd))

6. <u>Performance-Centered</u>: Use action verbs to begin each objective.

7. <u>Performance-Centered</u>: Make sure the action verb is concrete and suggests something measurable (e.g., use „Identify„ or „Enumerate„ or „Describe„ instead of „Learn„ or „Know how to„ „familiar„ „Explore„ or „Awareness„). Consider, for example, what ***is*** „familiar„?

8. <u>Performance-Centered</u>: Concretely identify behaviors that you expect to change. Remember that even when you are to assess affective learning objectives, a student's oral report of feelings, values, or interests during an interview or responses to a survey concerning interests are in the end, **behaviors.**

e. Learning Objectives may be classified as either Mastery or Developmental. Mastery learning objectives are called "can do" objectives because they focus on performances, specific feats or behaviors that do not develop over time such as square rooting a number, listing the functions of a cell, describing the Declaration of Independence. A developmental objective, on the other hand, is a statement that represents a broad domain of skills and/or abilities that are continuously developed throughout life, developed continuously to higher levels (rather than representing an all or none dichotomy implied y mastery objectives). Examples include objectives focusing on Writing, Reading, Problem Solving, etc.

f. Whereas an **<u>educational goal</u>** amounts to a general intention of instruction, each objective should be clear, specific, focused, and measurable. Objectives are written to operationalize the abstract intention of a goal. By explicitly expressing what a teacher specifically intends to do to attain a goal, objectives represent a concrete commitment to a course of action. Writing objectives can be difficult because the process forces the teachers to think through exactly what student benefits are expected.

Writing useful and meaningful objectives can require much labor and time. The biggest challenge before professionals who are writing objectives is ensuring that some set of objectives accurately and fully constitute a goal. In practice, no set of objectives will fulfill such an expectation, and so they should be continually revisited, studied, and refined.

2. Taxonomies of Learning Objectives (Cognitive, Affective, Psychomotor)

a. Bloom's Cognitive Taxonomy

1) Knowledge: recall of factual material in a similar form to that in which it was presented during instruction

2) Comprehension: translation interpretation or extrapolation of a concept into somewhat different form than originally practiced or presented.

3) Application: solving new problems through the use of familiar principles or generalizations

4) Analysis: breaking down a communication or problem into its component elements by using a process that requires recognition of multiple elements, relationships among these elements, and/or organizational principles.

5) Synthesis: combing elements into a whole by using an original structure or solving a problem that requires a combination of several principles sequentially into a novel situation

6) Evaluation: employment of internal (self-generated) or external criteria for making critical judgments in terms of accuracy, consistency of logic, or artistic or philosophical point of view

b. Dimensions of Learning Model

1) Declarative knowledge: the facts, ideals, generalizations, and/or theories to be assessed.

2) Procedural knowledge: the skills or procedures to be assessed.

3) Complex thinking: types of reasoning strategies and ways of applying knowledge.

4) Information processing: aspects of information gathering, synthesizing, evaluating, and needs assessment.

5) Effective communication: aspects of ideal communication, audience communication, purpose for communication, and products for communication.

6) Collaboration and cooperation: types of work on group goals, interpersonal skills, group maintenance activities, and multiple role activities.

7) Habits of mind: types of self-regulation, critical thinking and creative thinking performances.

c. Krathwohl's taxonomy of affective objectives

1) **Receiving** (attending): Being conscious of something, willingly give it attention, and controlling the fixation of one's attention on something despite competing and distracting stimuli.

2) **Responding**: Obedience to authority; voluntarily responding to instruction for reasons beyond fear, and deriving satisfaction or pleasure from such compliance.

3) **Valuing**: Emotional acceptance of a proposition or doctrine that one indeed considers intellectually tenable; Pursuing, seeking, and wanting a value; and a commitment to a value that inspires one to extend the possibility of that value and deepen one's involvement with the value.

4) **Value Organization**: The degree to which one conceptualizes how a value relates to other values already held, perhaps bringing multiple values together in an orderly system that depicts how the values relate to one another.

5) **Characterization by a value or Value Complex**: when a person integrates values and attitudes into a system that permits the individual to act consistently or behave in a principled manner. This may include the development of one's view of the universe or philosophy of life.

d. Harrow's taxonomy of psychomotor objectives

1) Reflex movements
2) Basic-fundamental movements
3) Perceptual abilities
4) Physical abilities (Endurance, Strength, Flexibility, and Agility)
5) Skilled movements
6) Non-discursive communication such as expressive movement or interpretative movement

e. Evaluating Learning Objectives

Learning Objectives ideally should be

1) Appropriate for the educational level of the students
2) Limited only to the important outcomes for the course
3) Consistent with the stateÂ published learning standards
4) Consistent with the local schoolÂ philosophy and general goals
5) Can be defended by currently accepted learning principles
6) Taught in the time limits of the course
7) Taught with available teaching resources

f. Making sure assessment tasks match learning objectives.

1) A very basic requirement for the validity of a classroom assessment procedure is that the procedures should match the intentions of the specific learning objectives in your assessment plan.

2) Because **developmental** learning objectives tend to be broad, more than one assessment procedure should be used to assess the objective so that the assessment results are valid and reliable, two concepts to be discussed in later classes

C. Validity of Assessment Results (cf. chapter 5)

1. Introduction to Validity

a. Cronbach described **validation** as the process by which a test developer or test user collects evidence to support the types of inferences that are to be drawn from the test scores. Said otherwise, **validity** is the soundness of your interpretations and uses of studentÂ assessment results. To plan a validation study, the desired inference must be clearly identified. Then an empirical study is designed to gather evidence of the usefulness of scores for such inferences.

b. The concept of validity applies to the ways in which we interpret and use the assessment results and not the assessment procedure itself. So properly stated, we ask, , Is it valid to interpret the scores from this test as measuring reading comprehension?„ rather than , Is this test valid?„

c. The assessment results have different degrees of validity for different purposes and for different situations. Tests that are crafted after learning objectives are going to be more valid for the situation at hand than other tests that attempt to address the same general area, that are more global in perspective, perhaps designed with a different use or philosophy in mind.

d. Judgements about the validity of your interpretations or uses of assessment results should be made only after you have studied and combined several types of validity evidence.

e. The interpretations (meanings) you give to your studentsÂassessment results are valid only to the degree that you can point to **evidence** that supports their appropriateness and correctness.

f. The uses you make of your assessment results are valid only to the degree to which you can point to **evidence** that supports their appropriateness and correctness.

g. The interpretations and uses you make of your assessment results are valid only when the **values** implied by them are appropriate.

h. The interpretations and uses you make of your assessment results are valid only when the **consequences** of these interpretations and uses are consistent with appropriate values.

2. Validity of Teacher-Made Classroom Assessment Results

Review the criteria for improving the validity of scores from classroom assessments used for assigning grades to students

Ask yourself following questions:

1. Does my assessment procedure emphasize what I have taught?
2. Do my assessment tasks accurately represent the outcomes specified in my schoolÂ or stateÂ curriculum framework?
3. Are my assessment tasks in line with the current thinking about what should be taught and how it should be assessed?
4. Is the content in my assessment procedure important and worth learning
5. Do the tasks on my assessment instrument require students to use important thinking skills and processes?
6. Does my assessment instrument represent the kinds of thinking skills that my schoolÂ or stateÂ curriculum framework and performance standards state are important?
7. Do students actually use the types of thinking I expect them to use on the assessment?
8. Did I allow enough time for students to demonstrate the type of thinking I was trying to assess?
9. Is the pattern of results in the class consistent with what I expected based on my other assessments of them?
10. Did I make the assessment tasks too difficult or too easy for my students?
11. Do I use a systematic procedure for obtaining quality ratings from student performances on the assessment?
12. Does my assessment instrument contain enough tasks relative to the types of learning outcomes I am assessing?
13. Do you word the problems or tasks on your assessment so students with different ethnic and socioeconomic backgrounds will interpret them in appropriate ways?
14. Did you modify the wording or the administrative conditions of the assessment tasks to accommodate students with disabilities or special learning problems?
15. Do the pictures, stories verbal statements, or other aspects of my procedure perpetuate racial, ethnic, or gender stereotypes?
16. Is the assessment relatively easy for me to construct and not too cumbersome to use to evaluate students?
17. Is the time needed to use this assessment procedure better spent on directly teaching students instead?
18. Does your assessment procedure represent the best use of your time?
19. Are the assessment results used in conjunction with other assessment results?

3. Validity of Extra-Classroom Assessments

a. Extra classroom assessments include district- and state-mandated assessments, standardized achievement and aptitude tests, attitude inventories, and individually administered intelligence tests.

b. There are Eight Categories of Validity Evidence

1. **Content Evidence**: Comes from judging the content of the tasks or items on the instrument in terms of their content relevance and curricular relevance. Also called CONTENT VALIDITY

   i Content representativeness: concerns the degree to which the assessment tasks or items are a representative **sample** of the larger **domain** of preference.

   ii Table of specifications (Test Blueprint; Test Specifications): A table used to define the domain for tests, surveys, performance tasks, etc. This table, also known as a Test Blueprint or Test Specifications, serves as an organizer that frames the major content categories and skills to be assessed. The proportion of the tasks or items that will be included on the instrument or overall performance should correspond roughly with how important the domain is relative to other domains. One way of gauging the importance of a domain is by considering how much time you spend on a topic during instruction.

Example #1

Elementary reading skills

Content Based Category	24	Comprehension	Application	Synthesis
Main Idea	6	3	2	1
Author's Viewpoint	3	3		
Inference	2		2	
Prediction	4		2	2
Thinking Maps	2		1	1
Summarizing	4	4		
Sequencing	3		3	

Example #2

Poetry and the Romatics: Test Specifications using Bloom's Taxonomy

Content Base Category	**35**	**Knowledge**	**Comprehension**	**Application**	**Analysis**	**Synthesis**	**Evaluation**
Vocabulary	**5**		5				
Historical Settings	**3**	2	1				
Interpretation	**7**			1	3	2	1
Rhyme	**3**	1			1		1
Meter	**3**	1		2			
Denotation & Connotation	**5**			1	2	2	
Symbolism	**5**		1	3	1		
Metaphor & Simile	**4**				2	2	

Example #3

Multicultural Awareness: Blueprint Table

Content Base Category	Number of items
Discrimination/Racism	7
Cultural differences	7
Lack of Visible Culture	5
Under Representation	5
Degree of ,Fit,,	5
Leadership skills	5
Study skills	5
Role Models	5
Language Barriers	5
Social Interactions	5
Recruitment of Students	4
Retention of Students	4
Financial Concerns	4
Family Issues	4
Hate Crimes	4

Example #4

Physical Education Sportsmanship Instrument: Blueprint Table

Content Base Category	34	Comprehension	Application	Evaluation
Define sportsmanship	16	3	5	8
Relate sportsmanship to game situations	5	5		
Appropriately model sportsmanship (team captains)	7		5	2
Articulate the value of sportsmanship	6	6		

Example #5

Cadet Supervisors: Table of Specifications using Bloom's Taxonomy

Content Base Category	28	Knowledge	Application	Evaluation
Crowd Control	2		2	
Parking Vehicles	5	3	1	1
Supervisory Training	5	5		
Special Event Organization	5		2	3
Shift Assignments	2	2		
Disciplinary Skills	4		2	2
Interpersonal Skills	5	1	2	2

3. Validity of Extra-Classroom Assessments

b. There are Eight Categories of Validity Evidence

1. Content Evidence (Cont'd)

i Content relevance: concerns whether the assessment tasks or items are included in the test userÂ domain definition.

ii Curricular relevance: concerns the degree of overlap between the school curriculum and assessment tasks. It is important that an assessment method is relevant to the schoolÂ definition of the achievement domain to the extent that it matches the schoolÂ curriculum learning objectives. It is important that the school also weights the content areas to the same extent as the assessment method.

2. **Substantive Evidence**: Concerns judging what kinds of thinking processes and skills students must use to complete the tasks/items successfully.

3. **Internal Structure Evidence**: The interrelationships among the tasks and between the tasks and the total results. Evidence from research studies that examine the Internal structure of the test/ performance should establish that the test/performance measures only as many domains as are intended. For example, if a test is created to measure arithmetic ability, then the test should measure only one domain, each item should relate sufficiently to the total test score or underlying construct. These interrelations are usually established by way of correlation coefficients or factor analysis (more on correlation coefficients later).

i. Factor analysis: a statistical procedure used to inspect the degree to which test items are correlated with the same construct. Each item is evaluated in its own right on the basis of its correlation with the construct. A higher correlation indicates that an item is very related to the construct. Items with lower correlations are considered for removal.

4. **External Structure Evidence**: Concerns the extent to which measure/performance assessment results relate to other assessment results obtain by other established tests that have been proven to address the same/similar content area. For example, scores obtained from a new spelling test should relate to scores obtained from a well-known and researched spelling test. Scores obtained from a new

3. Validity of Extra-Classroom Assessments

External Structure Evidence continued:

achievement test battery should relate well to scores obtained from the K-ABC or the SAT.

b. There are Eight Categories of Validity Evidence (Cont'd)

i. **Correlation with other established tests**. The correlation must not be too high nor too low. If the correlation is too high then the test may not be said to make a unique contribution (why not use the older test?).

ii. **Age differentiation**: Scores should get higher as age increases if the test measures a construct that is age related (i.e., score on an intelligence test)

5. **Reliability Evidence**: The consistency of the assessment results. (More on this later). You can not make valid interpretations of assessment results unless you have an assessment method (test or performance task) that gives scores/ratings consistently across time, across judges, across content domains. Reliability is a necessary but not sufficient condition for validity.

6. **Generalization Evidence**: Concerns how broadly one may interpret and use assessment results. Often the validity of our interpretations and uses of a test/performance exercise is limited to certain conditions.

7. **Consequential Evidence**: Concerns whether the intended consequences of a test/performance are attained for instruction learning, and equity.

8. **Practicality Evidence**: concerns whether there are practical barriers that impede the proper use of assessment results.

c. **The Correlation Coefficient**: MOST VALIDITY PROCEDURES described above requires the use of the correlation coefficient, at minimum. A correlation coefficient is a numerical value ranging from ‡ 1.0 to 1.0 that indicates the degree of relationship between two variables. A zero correlation suggests that there is no empirical evidence of a relationship between the two variables studied. As a value approaches 1.0 or ‡1.0, the relationship between two variables is considered to be higher. Either a ‡1.0 or 1.0 correlation suggests a perfect relationship

between the test scores, performance scores, etc. Negative correlations are interpreted differently than positive correlations. Positive correlations suggest that as scores increase on one test, score on the other test tend to increase also. For example, a positive correlation exists between the height and weight of a person. The taller someone is, the more that person tends to weigh. Negative correlations suggest that as score on one measure decrease in value, scores on some other measure tend to increase. For example, the slower a person drives a car, the more likely the person will not have an accident. As inflation increases, buying goes down. The less a city enforces the law, the more a city has criminal incidences.

d. Remember that you must be prepared to review and combine several types of evidence before judging the validity of a test/performance task.

D. Reliability of Assessment Results (cf. chapter 4)

1. General information

a. A point of clarity. The following terms are use interchangeably throughout the literature and your text: Test, instrument, measurement method, measurement procedure, exam, method of measurement, measure, measurement tool, assessment procedure, assessment tool, and assessment method, performance method, performance task.

b. Whenever a test is administered, the test administrator would like some assurance that the test results could be replicated if the same individuals were tested again under the same circumstances.

c. **Reliability**: The consistency (or reproducibility) of test scores. This consistency may be expected to occur when the same people (1) are reexamined with the same test on different occasions, or (2) receive two different forms of the a test on the same occasion, or (3) receive one form of a test on the same occasion. In the latter case, you want to know how consistently examinees were in responding to all items on the test (item homogeneity). The theory behind this is that the more consistent the examinees are in responding across items, the more consistent their performance is likely to be with future administrations.

d. **IMPORTANT**: Note that Reliability is a property of the assessment results rather than the property of the instrument. It is more appropriate to speak of the reliability of assessment scores rather than the reliability of an assessment, per se.

2. Procedures for Estimating Reliability

a. Reliability procedures calculated with **Two test administrations**

i. When you took the SAT, it should have been administered under controlled conditions at a particular site on a given date. Because cheating on the exam must be controlled, examinees in adjacent seats should have taken different forms of the exam covering the same content. The question is just how fair was it to give two different forms of a test to the examinees? Did one group receive an easier exam? or a more understandable exam? One way to answer this question is to use **Alternate Form reliability.**

2. Procedures for Estimating Reliability (Cont'd)

a. Reliability procedures calculated with **Two test administrations**

ii. **Alternate Form reliability**: indicates how consistently examinees respond to two similar forms of a test (different items, similar content). The two test forms are administered one right after the other to the same group of examinees (giving a break is OK to guard against burn-out). Then, a correlation coefficient is calculated between the scores obtained for each test form. The result is called a , ***coefficient of equivalence***„.

iii. **Test Retest reliability**: indicates how consistently the same examinees respond to a test over time. Calculate a Correlation coefficient between two administrations of the same test and call the result a , ***coefficient of stability***„. The problem with this type of reliability is that exposure to the test contents promotes better performances on later administrations of the same test (i.e., , practice makes perfect„). Moreover, if the test administrations are separated in time long enough so that the examinees forget the test contents, a new problem arises: maturation and outside learning most likely will occur and thus influence future test performance. A critical question in the design of a test-retest reliability study is this: How much time should elapse between testings? There is no single answer. The time should be long enough to allow effects of memory or practice to fade but not so long as to allow maturational or historical changes to occur in examineeÂ true scores. The purpose for which the test scores are to be used should be taken into account in designating the waiting time.

iv. **Test-Retest with alternative forms**: In this case, you administer one form of the test, wait for some specified period, then administer the other form of the test. Such reliability coefficients tend to be smaller in value than other reliability coefficients. The correlation coefficient measuring the relationship between the two forms is referred to as a ***coefficient of stability and equivalence***.

2. Procedures for Estimating Reliability(Cont'd)

b. Reliability procedures calculated with only **<u>One test administration</u>**

i. **<u>Split-Half reliability</u>**: Indicates how much of a relationship exists between two halves of a test. You can split a test into two halves in one of four ways: (1) You may randomly assign items to two groups; (2) you may assign all even numbered items to one group and odd numbered items to the other group; (3) you may rank order items according to difficulty levels based on the responses of examinees and then assign odd and even numbered items to two groups; or (4) you may match items according to content and then assign items similar content to different groups. After splitting the items into two groups, you calculate a correlation coefficient between scores for the two halves. The resulting value is called a ,coefficient of equivalence„. The problem with this ,coefficient of equivalence„ is that reliability will be underestimated, smaller than it should be. You **must** correct for the underestimation of reliability caused by splitting the test into two forms. To do so, you would plug the coefficient of equivalence into the **<u>Spearman Brown formula.</u>** The problem with this method is that different estimates are possible depending upon the way you split the test. Another problem is that this technique requires the two halves to be equivalent.

ii. **<u>Kuder-Richardson reliability procedures</u>** (i.e., KR 20 or KR21): Both procedures determine how internally consistent the items are and do not require a split between test halves. These procedures are appropriate when used for dichotomously scored items (i.e., when you have right-wrong answers).

iii. **<u>Cronbach's Alpha</u>** (i.e., Coefficient Alpha or simply „α„, the Greek letter for alpha): this procedure also determines how internally consistent the items are and does not require a split between test halves. This procedure is different from the Kuder-Richardson reliability procedures because it is not restricted to right-wrong responses, but may be used for polytomous responses as well (For example, when partial credit is given or survey responses with a Likert scale are examined). The CronbachÂ alpha represents the most general case for internal consistency, and so it may be used instead of the Kuder-Richardson or the coefficient of equivalence corrected by the Spearman-Brown.

2. Procedures for Estimating Reliability (Cont'd)

c. **<u>Interrater reliability</u>**: For some types of instruments only one set of items is used (a list of behaviors on a behavioral checklist), but multiple observations are collected for each examinee by having two or more raters complete the instrument. In this case, the consistency of the observations over raters may be of interest. The ratersÂ responses are correlated and hopefully the correlations are high and positive in value (say, .80 or higher). This would suggest that the raters score performances or products in a similar manner.

d. Factors that affect reliability coefficients

i. Longer assessment procedures tend to be more reliable

ii. The numerical value calculated for a reliability coefficient will fluctuate from one sample of persons to another.

iii. The narrower the range of a groupÂ ability, the lower the reliability coefficient tends to be.

iv. Students at different achievement levels may be assessed with different degrees of accuracy.

v. The longer the time interval between testings, the lower will test-retest and alternative forms reliability coefficients tend to be.

vi. More objectively scored assessment results are more reliable

vii. Often different methods of estimating reliability will not give the same result.

Part II Crafting and Using Classroom Assessments

A. Planning for Integrating Assessment and Instruction (cf. Chapters 7 & 11)

1. Purposes of Classroom Assessment

a. Formative uses: help teachers to monitor or guide student learning while it is still in progress

i. Sizing the students up before instruction begins
ii. Diagnosing individual learning needs (adapting instruction)
iii. Diagnosing group learning needs (reinforcing and re-teaching)
iv. Planning later instruction

b. Summative uses: help a teacher to evaluate student learning after teaching one or more units of a course of study.

i. Assigning grades for report cards
ii. Placing students into remedial and advanced courses
iii. Evaluating OneÂ's own teaching

c. Other uses: help in teaching generally but may not be linked to evaluating individuals

i. Assessments as teaching tools
ii. ControllingÂ studentsÂ behavior
iii. Communicating achievement expectations to students.

2. Plans for teaching and assessment should be developed together. You need to align what and how you teach with what and how you assess. You teach so students can achieve certain learning targets; you assess those targets to see, if in fact, students have achieved them.

Plans for a marking period usually apply to two or three units of instruction. A unit of instruction is a teaching sequence covering from one to seven weeks of lessons, depending on the students and the topics you are teaching.

B. Completion, Constructed-Response, and True-False Items (cf. Chapter 8 & 9)

1. Three fundamental principles for crafting assessments

a. Your assessment should focus on the important learning objectives.
b. Elicit from the students only their knowledge and performances relevant to the learning objective being assessed.
c. Neither prevents nor inhibits as studentÂ ability to demonstrate achievement of the learning objectives

2. Short answer and completion items

a. Short answer items require a student to respond to each item with a word, short phrase, number or symbol

b. Three kinds of short answer items exist

i. Short answer questions: ask a direct question

What is the capital city of Florida?

ii. Short answer completion items: requires a student to add words to complete an incomplete statement (Avoid using short answer completion items)

The capital city of Florida is _____

iii. Short answer association items: consists of a list of terms or a picture for which the student as have to recall numbers, labels, symbols, or other terms

On the blank next to the name of each chemical element, write the symbol used for it.

Element	***Symbol***
Barium	_____
Calcium	_____
Chlorine	_____
Potassium	_____
Zinc	_____

c. Strengths of short answer items

i. Short answer formats can be used to assess either low level or high level abilities

ii. Relatively easy to construct an can be scored more objectively

iii. Students have a low probability of getting the answer right by random guessing.

iv. Partial credit may be awarded for partial understandings

d. Shortcomings of short answer items

i. They are not totally free from subjective scoring (you cannot anticipate all of the responses student as will make). This affects the reliability of your scores and therefore the validity of your interpretations and uses.

ii. Spelling errors, grammatical errors, and legibility tend to complicate the scoring process further. This affects the reliability of your scores and therefore the validity of your interpretations and uses.

e. Principles to consult when crafting Short Answer and Completion items

1) Does the item assess an important aspect of the unitÂ instructional objectives?

2) Does the item match your assessment plan in terms of performance, emphasis, and number of points?

3) If possible, is the item written in question form thereby focusing the item on the specific knowledge sought?

4) Is the item worded clearly so that the answer is a brief phrase, single word, or single number?

5) Is the blank or answer space towards the end of the sentence?

6) Is the statement paraphrased rather than copied verbatim from the learning materials?

7) Is the word omitted in a completion item an important word rather than a trivial word?

8) Are there only one or two blanks?

9) Is the blank in this item: (a) the same length as the blanks in the other items; if appropriate, arranged in a column?

10) If appropriate, does the item (or directions) state the degree of detail, specificity, precision, or units you want the answer to have?

11) Does the item avoid grammatical (and other irrelevant) clues to the correct answer?

3. True-False Items

a. Different kinds

1. True- false items: present a proposition that is evaluated by student as being true or false.

Beijing is the capital of Alaska T F

---

2. Yes-no items: Ask a question, to which the answer is either yes or no.

Does it take three licks to get to the center of a Tootsie Roll Tootsie pop? Yes No

---

3. Right wrong items: present a proposition that is evaluated by student as being right or wrong.

$12 \text{ X } 3 = 6^2$ R W

---

4. Correction items: the student judges the truthfulness of the proposition, but is also required to correct any false statement to make it true.

Indicate whether the following sentences are correct or incorrect; if incorrect, change the underlined word or phrase to suit the sentence.

The new student, who we met today, came from Greece. **Correct Incorrect**

---

5. Multiple True-False items: similar to multiple choice items, however, every option is treated as a separate true-false statement (more than one choice may be true).

Under the Bill of Rights, freedom of the press means that newspapers:

1. Have the right to print anything they wish without restrictions. T F
2. Can be stopped from printing criticisms of the government. T F
3. Have the right to attend any meeting of the executive branch of the federal government. T F

---

3. True-False Items

a. Different kinds (Cont'd)

6. The yes no with explanation: requires students to answer yes or no to a question and then explain why her or his answer is correct.

If I want to increase the validity of a test, I may elect to increase the number of items. Am I correct? Yes No

If answer no, explain why this is wrong?

b. Strengths of True-False Items

i. Certain aspects of a subject matter readily lend themselves to verbal propositions that can be judged as true or false
ii. relatively easy to write
iii. scored easily and objectively
iv. can cover a wide range of content within a relatively short period
v. Well-written True-False items can assess more than simple recall

c. Short comings of True-False Items

i. assess only specific, frequently trivial facts
ii. can be ambiguously worded
iii. are susceptible to random guessing
iv. may encourage students to study and accept only oversimplified statements of truth and factual details

d. Principles to consult when crafting True-False items

1) Does the item assess an important aspect of the unitÂ instructional objectives?

2) Does the item match your assessment plan in terms of performance, emphasis, and number of points?

3) Does the item assess important ideas, knowledge or understanding (rather than trivial, general knowledge, or common sense)?

4) Is the statement either definitely true or false without adding further qualifications or conditions?

5) Is the statement paraphrased rather than copied verbatim from the learning materials?

6) Are the word lengths of true statements about the same as those of the false statements?

7) Did you avoid presenting items in a repetitive or easily learned pattern (e.g., TTFFTT÷ , TFTFTF÷)?

8) Is the item free of verbal clues that give away the answer? (i.e., **specific determiners** such as , always„, , never„, , every„ to make propositions **false**; , often„, , usually„, , frequently„, to make a proposition **true**.) A **specific determiner** is a word or phrase that , over-qualifies„ a given statement and gives the student an unintended clue to the correct answer.

9) If the statement represents an opinion, have you stated the source of the opinion?

10) If the statement does not assess knowledge between two ideas, does it focus on only one important idea?

C. Multiple Choice and Matching Exercises (cf. Chapter 8 & 9)

1. Key terms

a. Multiple choice items: consist of one or more introductory sentences followed by a list of two or more suggested responses.

b. Stem: the part of the item that asks the question or introduces an incomplete sentence.

c. Alternatives: The test takers presented list of suggested responses on a multiple-choice item (also known as choices, responses, and options).

d. Keyed answer: The correct alternatives.

e. Distractors: The remaining incorrect alternatives (also known commonly as foils). Distractors are meant to be plausible (but incorrect) answers to the question (or solutions to the problem) in the stem. That is, distractors ought to be plausible to those students who have not sufficiently mastered the material.

f. Interpretative material: Information presented before the introduction of the item in order to make an item more authentic/ clear or relevant. This preliminary information might set the stage for one or more items to prepare the test taker with a stimulus or foundation for the upcoming content of an item. Interpretative materials may include tables, charts, graphs, diagrams, pictures or scenarios.

g. Context-dependent items: Items based upon initially presented interpretative material. They are also called interpretative exercises or linked items.

C. Multiple Choice and Matching Exercises (Cont'd)

2. Five kinds of Multiple choice items

a. Correct Answer

According to BloomÂ taxonomy, what is Knowledge?

A. The recall of factual material in a manner similar to how it was originally presented
B. The specialized culinary tool Sam uses to stir fry green eggs and ham.
C. A magnificent yellow crocus opened in a verdant prairie and subject to an April IRS audit.
D. Stuff a person knows, located in the cerebella quarters, you know, between two ears.

b. Best answer

According to Nitko, how should short answer items ordinarily be crafted? As a(n)

A. Completed response.
B. Question.
C. Multiple choice item.
D. Essay item

c. Multiple response

Which use(s) of assessment have been identified by Nitko as summative? (Circle all that apply.)

A. Planning later instruction
B. Assigning grades for report cards
C. Placing students into remedial and advanced courses
D. Diagnosing group learning needs

d. Incomplete statement

The reliability procedure that may be used to obtain the internal consistency of nondichotomous items is:

A) Kuder-Richardson
B) Alternate Form method
C) CronbachÂ alpha
D) Test Retest method

C. Multiple Choice and Matching Exercises (Cont'd)

e. Negative (Avoid using these)

Which of the following is NOT a shortcoming for **true-false** items?
A. assess only specific, frequently trivial facts
B. are susceptible to random guessing
C. spelling errors tend to complicate scoring
D. can be ambiguously worded

3. Advantages of Multiple choice items

a. Can be used to assess a greater variety of learning objectives than other formats of response choice items

b. Do not require students to write out or elaborate their answers

c. Minimize the possibility that students will dress-up or bluff their answers with verbiage.

d. Focus on reading and thinking.

e. A review of distractors permits teachers to diagnosis the difficulties that students are experiencing with the item content.

4. Criticisms of Multiple choice items

a. Students must choose from a fixed set of options rather creating or expressing their own ideas or solutions.

b. Poorly written items can be superficial, trivial, or limited to factual knowledge.

c. Because usually only one option of an item is keyed as correct, brighter students may be penalized for not choosing it due to flaws, ambiguities in wording, or divergent viewpoints.

d. Multiple-choice items tend to be based on standardized, vulgarized, or approved knowledge giving student the impression that only one correct answer exists for any problem in a given area.

e. Exclusive use of Multiple-choice items in high stakes testing for important or high-stakes assessments may shape education in undesirable ways. The standard teacher response of offering drill and practice techniques may be inappropriate for items that assess using knowledge and applying higher order thinking skills.

5. Principles to consult when crafting Multiple choice items

a. Correct choice should about the same length as the distracters/foils (i.e., incorrect alternatives).

b. Correct choice should be different from distracters in meaning only, with no superficial verbal clues (specific determiners). No grammatical cues.

c. If an item depends in any way upon another, neither should reveal the answer to the other.

d. Answers should follow a random pattern.

e. There should be a clear central problem in the stem of each item. Use a question as the stem if possible.

f. Wording should be as concise as possible.

g. Information in the stem should be complete enough to make one answer justifiable. There should sufficient specificity in the item.

h. Distracters should be plausible. This is this most important criterion.

i. Vocabulary should be appropriate to the group for which the test is intended.

j. Points tested should be relevant, not trivial.

k. Avoid confusing sentence structure.

l. Vary homogeneity of alternatives to attain desired difficulty level.

m. Understanding of terms is better tested by placing term in stem and alternative definitions in options rather than by placing definition in stem and terms in options. In short, put the term in the stem.

n. State stem in positive form, if positive. Emphasize negative wording whenever it is used in the stem.

o. Items used to measure understanding should contain some novelty.

p. Use special alternatives such as , none of the above„ or , all of the above„ sparingly, if at all.

q. If alternatives are numerical, arrange them in order (either ascending or descending).

r. Use an efficient item format.

6. Matching Items

a. Terms

1) Premises: The initial column that contains numerically labeled terms, propositions, etc. A blank space is provided before each of the premises so that test takers can have a place to insert their answer.

2) Responses: The second column that contains alphabetically labeled terms, pictures, or other response options.

3) Perfect Matching is undesirable in the eyes of most assessment specialists because students are automatically credited for responses in which the answer was deduced by process of elimination alone.

b. Types of Matching

1) Masterlist

A. Interest assessments
B. Power assessments
C. Value assessments
D. Speed assessments

__ 1. Time limits are generous
__ 2. Performance speed is a key focus
__ 3. Preference for particular activities

2) Keylist (Classification)

___ 1. Time limits are generous
___ 2. Performance speed is a key focus
___ 3. Preference for particular activities

A. Interest assessments
B. Power assessments
C. Value assessments
D. Speed assessments

c. Advantages of Matching Items

Space saving, compact, and objective way to assess a number of important concepts, relating two sets of things.

d. Criticisms of Matching Items

1) Encourages Rote memorization
2) Tests Rote associations only
3) Finding homogeneous premises and responses can be difficult.

7. Matching Items (Cont'd)

e. Checklist for Matching Items

1) Within this exercise do the premises and responses all belong to the same category?

2) Do your directions clearly and completely explain the basis you intend students to use for matching?

3) Does every element in the response list function as a plausible alternative to every element in the premise list?

4) Are there fewer than 10 responses in this matching exercise?

5) Did you avoid perfect matching?

6) Are the longer statements in the premise list and the shorter statements in the response list?

7) If possible, are the elements in the response list ordered in a meaningful way (logically, numerically, alphabetically)?

8) Are the premises numbered and the response elements lettered?

D. Essay Assessment Tasks (cf. Chapter 9)

1. **Rater drift**: the tendency to change the way scoring criteria are applied over time.
2. **Halo effect**: our judgments of one characteristic of a person are influenced by our judgments of other characteristics or by our general impression of that person
3. **Carryover effect**: when your judgment of a students response to Question 1 affects your judgment of the students response to Question 2.

4. Checklist for Essay Assessment Tasks (a - j)

a. Does the item test an important aspect of this unit's learning objective?

b. Does the item match your table of specifications in terms of required performance, emphasis and number of points?

c. Does the item require student to apply their knowledge or skill to a new or novel situation?

d. When viewed in relation to other items on the test, does this item contribute to covering the range of content and behavior specified in your test plan?

e. Is the item focused? Does it define a task with specific directions, rather than leave the assignment so broad that virtually any response can satisfy the question?

f. Is the task defined by the item within the level of complexity that is appropriate for the educational maturity of the students?

g. To get a good mark on the item, is the student required to demonstrate more than recall of facts, ideas, lists, definitions, generalizations, etc.?

h. Is the item worded in a way that leads all students to interpret the assignment in the way you intended?

4. Checklist for Essay Assessment Tasks (Cont'd)

i. Does the wording of the item make clear to students the following:
   1. Magnitude or length of the required writing?
   2. Purpose for which they are writing?
   3. Amount of time to be devoted to answering this item?
   4. Basis on which their answers will be evaluated?

j. If the essay item asks students to state and support their opinions on controversial matters, does the wording of the item clearly indicate that the students' assessment will be based on the logic and evidence supporting their arguments, rather than on the actual position taken or opinion stated?

5. **Scoring Essay Assessments**

a. **Analytic scoring rubrics**: an outline or list of the major elements that students should include in the ideal answer.

b. **Holistic scoring rubrics**: making a judgment about the overall quality of the student's response. (You may decide beforehand how many quality categories into which you will sort the student responses such as A, B, C, D, and F).

6. Holistic scoring rubrics

a. Advantages
1. You can score the students' papers a little faster than with analytic rubrics
2. It helps you to view the papers as a working whole.

b. Disadvantages
1. You give a single overall mark and do not point out the details to your students that might help them improve.
2. Your own bias (e.g., toward neatness or correct spelling) and errors (e.g., paying more attention to the correctness of a specific element in one student's papers than to another student) can be easily masked by an overall mark.

7. Analytic scoring rubrics

a. Advantages

1. By scoring each element separately, you can give students feedback as to their strengths and weaknesses.
2. By scoring each part separately, you can look over all the papers to see which elements of the answer gave students the most trouble and therefore need to be retaught.
3. By weighing some elements of the answer more heavily than others, you must face up to your own values (i.e., you must decide which elements you value more than others).

b. Disadvantages

1 Your scoring will be a little slower with an analytic scoring rubric.
2 For some essays, you may find it difficult to come up with well-defined elements in the scoring guide.
3 Beginning teachers may feel a bit frustrated by the amount of time needed to create a useful analytic scoring rubric.

8. Summary of principles for scoring responses to subject matter essays.

1. Prepare some type of scoring (e.g., an outline, a rubric, an "ideal answer", or specimen responses from past administrations)

2. Grade all responses to one question before moving on to the next question.

3. Periodically re-score previously scored papers

4. Score penmanship, general neatness, spelling, use of prescribed format, and English mechanics separately from the subject matter correctness.

5. Score papers without knowing the name of the pupil writing the response.

6. Provide pupils with feedback on the strengths and weaknesses of their responses.

7. When the grading decision is crucial, have two or more readers score the essays independently.

E. Performance, Alternative, and Authentic Assessments (Chapter 10)

1. **Performance Assessments**: presents a hands-on task to a student and uses clearly defined criteria to evaluate how well the student achieved the application specified by the learning objective. Requires students to apply their knowledge and skills from several areas to demonstrate that they can perform a learning objective.

2. **Performance task**: an assessment activity that requires a student to demonstrate achievement by producing an extended written or spoken answer, by engaging in group or individual activities, or by creating a specific product. Students directly demonstrate the learning objective unless they are only required to provide a brief response. Two aspect of a student's performance includes the **Product** and the **Process.**

3. **Alternative assessment**: another word for performance assessment used to contrast this kind of assessment from standardized achievement tests, multiple-choice items, etc.

4. **Authentic assessment**: used when a performance assessment reflects a realistic and meaningful activity for some context.

5. Howard Gardner's theory of multiple intelligences as a taxonomy for performance assessment test specs.

6. Types of Performance Assessment Techniques

a. Structured, On-demand tasks for Individual Students, Groups or Both (Paper and Pencil or Other kinds of tasks)

b. Naturally occurring or Typical Performance Tasks

c. Longer-Term Projects (for Individuals and/or Groups)

d. Portfolios (Best work Portfolios ‡i.e., summative- or Growth and learning progress Portfolios ‡ i.e., formative)

e. Demonstrations
f. Experiments
g. Oral Presentations, Debates and Dramatizations

h. Simulations and Contrived Situations (Realistic scenarios perhaps with actors or computerized audio-visual/test scenarios or simulations)

7. Advantages of Performance Assessments

a. Performance tasks clarify the meaning of complex learning objectives.
b. Performance tasks assess the ability "to do".
c. Performance assessment is consistent with modern learning theory (Constructivist).
d. Performance tasks require integration of knowledge, skills, and abilities.
e. Performance assessments may be linked more closely with teaching activities.
f. Performance tasks broaden the approach to student learning assessment.
g. Performance tasks let the teachers assess the processes students use as well as the products they produce.

8. Disadvantages of Performance Assessments

a. High-quality performance tasks are hard to craft.
b. High-quality scoring rubrics are hard to craft.
c. Completing performance tasks takes students a lot of time.
d. Scoring performance task responses takes a lot of time
e. Scores from performance tasks may have lower scorer reliability compared to multiple-choice and other objective items.
f. Student performance on one task provides little information about student performance on other tasks. (You may have to include several performance tasks to adequately measure one unit of instruction- the validity of your interpretations and uses are at stake).
g. Performance tasks do not assess all learning objectives well.
h. Completing performance tasks may be discouraging to less able students.
i. Performance assessments may underrepresent the learning of some cultural groups. (Performance tasks will not wash away cultural differences, but are more likely to make such differences more apparent. Assessors who are unaware of how different cultural groups express their higher thinking skills may be systematically biased in their assessments of them). Performance assessments may be corruptible.

F. Performance Tasks, Portfolios, Rating Scales, and Scoring Rubrics (Chapter 10)

1. **Performance Assessments**: presents a hands-on task to a student and uses clearly defined criteria to evaluate how well the student achieved the application specified by the learning objective. Requires students to apply their knowledge and skills from several areas to demonstrate that they can perform a learning objective

2. **Crafting Performance Tasks**

   a. Be very clear about the performance you want to assess.

   1) Select the learning objective(s) to assess.

   2) Specify the standards/quality dimensions (knowledge, skills and abilities) against which you will assess the students' performance. The standards can be content standards or lifelong standards. **Content standards** include specific declarative and procedural outcomes you want the student to achieve. **Declarative outcomes** are facts, ideas, generalizations, and theories you want the student to learn. **Procedural outcomes** are skills and procedures you want the students to learn.

   Lifelong standards include outcomes that cut across the curricula or may be used outside of the school, such as complex thinking, information processing, effective communications, cooperation, collaborations, and habits of mind.

   Each standard is referred to as a quality dimension. You should frame your performance task around them or some other framework that your school district requires.

   b. Limit dimensions assessed

   You should not try to assess all of the standards (i.e. quality dimensions) provided in class in one performance task, or the task will become unwieldy and confusing. Every performance task should assess one quality dimension from each of the following categories: content, complex thinking, information processing, and effective communication. Assessing one quality dimension from each of the other two categories (cooperation/ collaboration, and habits of mind) is optional.

c. Define the Quality Dimensions (Standards)

Each quality dimension you specify represents a continuum of educational growth. Different students will perform with different levels of quality on each dimension. Further, one student may perform with high competence on some dimensions but with less competence with others. Thus, part of crafting your performance task is to define the scale for each dimension. You define this quality scale by spelling out the different degrees of quality performance- from low to high - on each dimension. This continuum forms the basis for crafting scoring rubrics.

3. Crafting the task

a. Develop the Task in Nine Stages.

Marzano, Pickering, and McTighe (1993) identify 9 steps useful for developing a performance task.

1) Select a content dimension to build your task around (i.e. declarative or procedural knowledge)

2) Using this content dimension as a guide, select one of the complex thinking dimensions from Appendix E. that is closely related to the content dimension. These two dimensions will be the main focus of your task.

3) Using the content and thinking dimensions, draft your performance task. Craft the task so that students know they are required to apply the appropriate thinking skills standards to the content.

4) Select one appropriate information-processing dimension consistent with your content and thinking skills dimensions and with the task you are crafting.

5) (Optional) If your task is a group task, select a collaboration/cooperation (or a , habits of mind„) dimension to assess in conjunction with dimensions already selected.

6) Rewrite your performance task if you decided to use one or more of the standards described in Step 5.

7) Select an effective communication dimension you believe is important to assess with this task.

8) Rewrite the performance task to incorporate the effective communication dimension.

9) Review and edit the task. For each dimension, specify several quality levels of performance competence.

b. Checklist for judging the quality of performance tasks

1) Does the task focus on an important aspect of the unit's learning objectives?

2) Does the task match your assessment plan in terms of performance, emphasis, and number of points?

3) Does the task require the student to actually DO something rather than only write about how to do it or recall or copy information?

4) Do you allow enough time so that all of the students can complete the task under your conditions?

5) If this is an open-response task, do your wording and directions make it clear to students that they may use a variety of approaches and strategies, that you will accept more than one answer as correct, and that they need to fully elaborate their response?

6) If the task is intended to be authentic or realistic, do you present a situation that your level of students will recognize as coming from the real world?

7) If the task requires using resources and locating information outside of the classroom, will all of your students have fair and equal access to the expected resources?

b. Checklist for judging the quality of performance tasks (Cont'd)

8) Do your directions and other wording:

i) define the task that is appropriate to the educational maturity of your students?

ii) lead all students, including those from diverse cultural and ethnic backgrounds, to interpret the task requirements in that way that you intend?

iii) make clear the purpose or goal of the task?

iv) make clear the length or degree of elaboration of the response you expect?

v) make clear the bases on which you will evaluate the responses to the task?

9) Are the drawings, graphs, diagrams, charts, manipulatives, and other task materials clearly drawn, properly constructed, appropriate to the intended performance, and in good working order?

10) Do you need to modify or adapt the task to accommodate students with disabilities?

4. Two ways of creating scoring rubrics.

The first way:

a. Adapt or create a conceptual framework of quality dimensions that describe the content and performance processes that you should use.

b. Develop a detailed outline that arranges the content and process from Step 1 in a way that identifies what you should include in the general rubric.

c. Craft a general scoring rubric that conforms to this detailed outline and focuses on the important aspects of content and process to be assessed across different tasks. The general rubric will be used to craft specific rubrics.

d. Craft a specific scoring rubric for a specific performance task.

e. Use the specific scoring rubric to assess the performances of several students; use this experience to revise the rubric as necessary.

The second way:

a. Obtain copies of about 10 to 12 students' actual responses to a performance item.

b. Read the responses and sort all of them into three groups: high, medium and low quality.

c. After sorting, carefully study each student response within the groups, and write very specific reasons you would put that student response into that group.

d. Look at your comments across all categories and identify the emerging dimensions.

e. Separately for each of the three quality dimensions levels, write as specific student centered description of what responses at that level are typically like.

5. Checklist for judging scoring rubrics, checklists, and rating scales

a. Overall, does the rubric emphasize the most important content and processes of the learning objective?

b. Will the scores you get from the parts of the rubric (standards) match the emphasis that you give them in your assessment plan?

c. Do the total number of marks obtained from the rubric match the emphasis given the learning objective?

d. Will your students understand the rubric?

e. Are the categories rated within rubric suitable for giving your students the guidance they need top improve their performance on the learning objective?

f. Is the rubric for this particular task a faithful application of the general rubric?

g. Are the levels of the scales for the parts of the rubric (standards) described clearly in terms of the performance you can observe the students doing?

h. Does the rubric allow you to assess the student's use of appropriate declarative and procedural content and processes?

i. If the purpose of the task is to assess student's use of alternative correct answers/products or alternate correct processes/strategies, does the rubric clearly describe how each is to be rated and marked?

j. Does the rubric allow you to distinguish a wide range of student quality levels of performance on this task rather than putting all students into one or two quality levels?

Professional Responsibilities, Ethical Behaviors, and Legal Requirements in Educational Assessments (cf. Chapter 17)

You have a professional, ethical, and legal responsibility concerning the way you craft, use, and report the results of your classroom assessments. Professional Associations have developed codes of ethics and professional responsibilities.

1. **Standards for Teacher Competence in Educational Assessment of Students** was jointly developed by the American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association. These standards are intended for use as

   - A guide for teacher educators as they design and approve programs for teacher programs
   - A self-assessment guide for teachers in identifying their needs for professional development in student assessment.
   - A guide for workshop instructors as they design professional development experiences for in-service teachers.
   - An impetus for educational measurement specialists and teacher trainers to conceptualize student assessment and teacher training in student assessment more broadly than has been the case in the past.

2. **The Code of Fair Testing Practices in Education** was jointly developed by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education.

   States the major obligations to test takers of professionals who develop or use educational tests. The Code is meant to apply broadly to the use of tests in education (admissions, educational assessment, educational diagnosis, and student placement). The Code is not designed to cover employment testing, licensure or certification testing, or other types of testing. Although the Code has relevance to many types of educational tests, it is directed primarily at professionally developed tests such as those sold by commercial test publishers or used in formally administered testing programs. The Code is not intended to cover tests made by individual teachers for use in their classrooms.

3. The **Code of Professional Responsibility in Educational Measurement** was developed by the National Council on Measurement in Education.

The purpose of this Code is to guide the conduct of NCME members who are involved in any assessment activity in education. NCME also provided the Code as a public service for all individuals who are engaged in educational assessment activities in the hope that these activities will be conducted in a professionally responsible manner. Persons who engage in these activities include local educators such as classroom teachers, principles and superintendents. Review the Code to identify other groups for whom this Code was written.