

EOP Learning Communities 2007

**Assessing Student Learning:
How Do We Know That They Know?**

***Dr. Mark Stoner,
Interim Director,
Center for Teaching and Learning***

***Dr. Phil Smith,
Professor of Mathematics***
President, Academic Senate, American River College
Chair, Los Rios District Curriculum Coordinating Committee

Table of Contents

Purposes of assessment	p.3
Aligning Objectives with Evaluations	p.3
Understanding Validity of Assessment	p.3
Types of Tests	p.4
Alternative Test Methods	p.6
Creating Grading Rubrics	p.7
Example: Research Paper Rubric to Evaluate Early/Rough Draft	p.7
Example: Research Paper Rubric to Evaluate Final Draft	p.9
Resources for Rubrics	p.9
Bloom's Taxonomy of Educational Objectives (Cognitive)	p.10
Sample Multiple-Choice Items That Measure at Various Levels	p.11
Single Concept Measured at Different Levels	p.12
Quick Bits on Testing	p.14
Other resources	p.15

The entire handout is available at: <http://wwwctl.csus.edu/resources.htm>

Purposes of assessment

Without doubt, our first concern with assessment is *knowing what students know*. We need this information for many purposes: grading students; choosing the next learning objectives; providing evidence of effectiveness of teaching and learning.

Knowing more about assessment also gives us insight into how we *shape our students' thinking* by our evaluations. We have much to do with how they study, what they learn and how they learn it by the types of assessments we use. If we want students to think in a variety of ways and learn different kinds of content in different ways, we need to have a repertoire of assessments.

At the same time, we make of assessments as *instructional tools*. Pre-tests can help students zero in on material to which they need to pay particular attention; post-test discussions can provide opportunity to re-teach important ideas; opportunities for revisions or second-tries on a limited or unlimited basis can help students rehabilitate or deepen their knowledge of important topics.

Aligning Objectives with Evaluations

Students often complain of being tested on material they feel was not treated in the prior segment of the course. Sometimes, they're right and such a disconnect exists. If we are to know what they know, it is important that we align our learning objectives with the evaluations used.

Example 1:

If we start with an objective such as, "Students will *define* the concept of 'reliability,' we know that it is a lower order objective [See Bloom's Taxonomy] that requires memory and correctness. A multiple-choice, matching or short-answer item that asks the students to identify or write the definition proper aligns goal and measure. Asking the students to *create* a definition of "validity" based on inferences they can make about it from the discussion of reliability is not appropriate.

Example 2:

If we start with an objective such as, "Students will devise a theory-driven model of authentic assessment in calculus," we know that this is a complex, higher order (synthetic) outcome that requires recall, understanding, interpretation, creativity, and problem-solving ability. An appropriate assessment could be an essay, an oral exam, a take-home or open book exam, depending on the memory demands we feel are needed to know what students can do. A multiple choice item or matching item would be inappropriate.

Understanding Validity of Assessment

Simply put, validity is the quality of measuring what we mean to measure. For example, weighing a person does not measure what they eat. It may be reliable since repeated weighings get the same result, but weight is not a valid measure.

Validity is not easy to ensure, but we can make a few fairly easy checks to improve it. One is to compare the objective with the measure as in our example above. A second check on validity to check scores against students' demonstrated abilities. In short, there may be a problem with validity if the strong students miss items and weak students

don't. If there are numerous problems with students understanding what items meant, the item is probably not valid. If students generally cannot finish the test, validity becomes questionable. Keep in mind, too, that you are sampling learning, not assessing all that students did or could have learned.

Finally, a helpful source of feedback is "peer review." Ask a colleague who teaches the same courses or material to look at your exam

Types of Tests

Multiple-choice tests. Multiple-choice items can be used to measure both simple knowledge and complex concepts. [For examples, see "Multiple-Choice Items That Measure at Various Levels" in this packet.] Since multiple-choice questions can be answered quickly, you can assess students' mastery of many topics on an hour exam. In addition, the items can be easily and reliably scored. However, good multiple-choice questions are difficult to write. For a tutorial, visit: http://web.utk.edu/~mccay/apdm/mchoice/mc_b.htm

True-false tests. Because random guessing will produce the correct answer half the time, true-false tests are less reliable than other types of exams. However, these items are appropriate for occasional use. Some faculty who use true-false questions add an "explain" column in which students write one or two sentences justifying their response. For more information [visit:http://www.studygs.net/tsttak2a.htm](http://www.studygs.net/tsttak2a.htm)

Matching tests. The matching format is an effective way to test students' recognition of the relationships between words and definitions, events and dates, categories and examples, and so on. For more information [visit:http://tlt.its.psu.edu/suggestions/questionwriting/match_construct.shtml](http://tlt.its.psu.edu/suggestions/questionwriting/match_construct.shtml)

Essay tests or assignments. Essay tests or assignments enable you to judge students' abilities to organize, integrate, interpret material, and express themselves in their own words. Research indicates that students study more efficiently for essay-type examinations than for selection (multiple-choice) tests: students preparing for essay tests focus on broad issues, general concepts, and interrelationships rather than on specific details, and this studying results in somewhat better student performance regardless of the type of exam they are given (McKeachie, 1986). Essays also give you an opportunity to comment on students' progress, the quality of their thinking, the depth of their understanding, and the difficulties they may be having. However, because essay tests pose only a few questions, their content validity may be low. In addition, the reliability of essay tests is compromised by subjectivity or inconsistencies in grading. For further discussion, visit: http://www.idea.ksu.edu/papers/Idea_Paper_17.pdf [Note: consistency of grading can be substantially enhanced using rubrics. See "Creating Grading Rubrics" in this packet.]

A *variation of an essay test* asks students to correct mock answers. One faculty member prepares a test that requires students to correct, expand, or refute mock essays. Two weeks before the exam date, he distributes ten to twelve essay questions, which he discusses with students in class. For the actual exam, he selects four of the questions and prepares well-written but intellectually flawed answers for the students to edit, correct, expand, and refute. The mock essays contain common misunderstandings,

correct but incomplete responses, or absurd notions; in some cases the answer has only one or two flaws. He reports that students seem to enjoy this type of test more than traditional examinations. Such a design also allows you to focus on specific elements of understanding which provides a bit more structure for students taking the exam and in grading it.

Short-answer tests. Depending on your objectives, short-answer questions can call for one or two sentences or a long paragraph. Short-answer tests are easier to write, though they take longer to score, than multiple-choice tests.

They also give you some opportunity to see how well students can express their thoughts, though they are not as useful as longer essay responses for this purpose. For further discussion, visit:

<http://www.utc.edu/Administration/WalkerTeachingResourceCenter/FacultyDevelopment/Assessment/test-questions.html#short%20answer>

Problem sets. In courses in mathematics and the sciences, your tests can include problem sets. As a rule of thumb, allow students ten minutes to solve a problem you can do in two minutes.

Oral exams. Oral exams are sometimes used for undergraduates in foreign language classes. In other classes they are usually seen as too time-consuming, too anxiety provoking for students, and too difficult to score unless the instructor tape-records the answers. However, a math professor has experimented with individual thirty-minute oral tests in a small seminar class. Students receive the questions in advance and are allowed to drop one of their choosing. During the oral exam, the professor probes students' level of understanding of the theory and principles behind the theorems. He reports that about eight students per day can be tested. For more information, visit: <http://delivery.acm.org/10.1145/1070000/1067487/p143-gharibyan.pdf?key1=1067487&key2=9592967711&coll=GUIDE&dl=GUIDE&CFID=21135640&CFTOKEN=53369760> (Sections 4.1 & 4.2)

Performance tests. Performance tests ask students to demonstrate proficiency in conducting an experiment, executing a series of steps in a reasonable amount of time, following instructions, creating drawings, manipulating materials or equipment, or reacting to real or simulated situations. Performance tests can be administered individually or in groups. Performance tests can be useful in classes that require students to demonstrate their skills (for example, health fields, the sciences, education).

If you use performance tests,

- Specify the criteria to be used for rating or scoring (for example, the level of accuracy in performing the steps in sequence or completing the task within a specified time limit).
- State the problem so that students know exactly what they are supposed to do (if possible, conditions of a performance test should mirror a real-life situation).
- Give students a chance to perform the task more than once or to perform several task samples.

Alternative Test Methods

Take-home tests. Take-home tests allow students to work at their own pace with access to books and materials. Take-home tests also permit longer and more involved questions, without sacrificing valuable class time for exams. Problem sets, short answers, and essays are the most appropriate kinds of take-home exams. Be wary, though, of designing a take-home exam that is too difficult or an exam that does not include limits on the number of words or time spent. Also, be sure to give students explicit instructions on what they can and cannot do: for example, are they allowed to talk to other students about their answers? A variation of a take-home test is to give the topics in advance but ask the students to write their answers in class. Some faculty hand out ten or twelve questions the week before an exam and announce that three of those questions will appear on the exam.

Open-book tests. Open-book tests simulate the situations professionals face every day, when they use resources to solve problems, prepare reports, or write memos. Open-book tests tend to be inappropriate in introductory courses in which facts must be learned or skills thoroughly mastered if the student is to progress to more complicated concepts and techniques in advanced courses. On an open-book test, students who are lacking basic knowledge may waste too much of their time consulting their references rather than writing. Open-book tests appear to reduce stress but research shows that students do not necessarily perform significantly better on open-book tests. Further, open-book tests seem to reduce students' motivation to study. A compromise between open- and closed-book testing is to let students bring an index card or one page of notes to the exam or to distribute appropriate reference material such as equations or formulas as part of the test.

Group exams. Some faculty have successfully experimented with group exams, either in class or as take-home projects. Faculty report that groups outperform individuals and that students respond positively to group. For example, for a fifty-minute in-class exam, use a multiple-choice test of about twenty to twenty-five items. For the first test, the groups can be randomly divided. Groups of three to five students seem to work best. For subsequent tests, you may want to assign students to groups in ways that minimize differences between group scores and balance talkative and quiet students. Or you might want to group students who are performing at or near the same level (based on students' performance on individual tests). Some faculty have students complete the test individually before meeting as a group. Others just let the groups discuss the test, item by item. In the first case, if the group score is higher than the individual score of any member, bonus points are added to each individual's score. In the second case, each student receives the score of the group. Faculty who use group exams offer the following tips:

- Ask students to discuss each question fully and weigh the merits of each answer rather than simply vote on an answer.
- If you assign problems, have each student work a problem and then compare results.
- If you want students to take the exam individually first, consider devoting two class periods to tests; one for individual work and the other for group.
- Show students the distribution of their scores as individuals and as groups; in most cases group scores will be higher than any single individual score.

A variation of this idea is to have students first work on an exam in groups outside of class. Students then complete the exam individually during class time and receive their own score. Some portion of the test items are derived from the group exam. The rest are new questions. Or let students know in advance you will be asking them to justify a few of their responses; this will keep students from blithely relying on their work group for all the answers.

Paired testing. For paired exams, pairs of students work on a single essay exam, and the two students turn in one paper. Some students may be reluctant to share a grade, but good students will most likely earn the same grade they would have working alone. Pairs can be self-selected or assigned. For example, pairing a student who is doing well in the course with one not doing well allows for some peer teaching. A variation is to have students work in teams but submit individual answer sheets.

Portfolios. A portfolio is not a specific test but rather a cumulative collection of a student's work. Students decide what examples to include that characterize their growth and accomplishment over the term. While most common in composition classes, portfolios are beginning to be used in other disciplines to provide a fuller picture of students' achievements. A student's portfolio might include sample papers (first drafts and revisions), journal entries, essay exams, and other work representative of the student's progress; it should included a reflective essay explaining effects, outcomes, uses of ideas, skills and values documented in the portfolio. You can assign portfolios a letter grade or a pass/not pass. If you do grade portfolios, you will need to establish clear criteria.

Adapted from: <http://honolulu.hawaii.edu/intranet/committees/FacDevCom/guidebk/teachtip/quizzes.htm>

Creating Grading Rubrics

What steps are involved in developing a rubric?

- Decide what criteria are used in the discipline to define "quality performance."
- Gather sample rubrics (not necessarily from the discipline) which can be adapted to your purposes
- Gather samples of work (could be from both students and experts) that illustrate a range of quality
- Discuss (with students and/or colleagues) the characteristics of the work that distinguish good from poor examples
- Write "objective" descriptors or definitions for each important characteristic. These describe what the trait is about, not what good performance looks like.

Objective:
Writing style was consistent.

Subjective:

Writing style really captured my interest.

- Describe strong, middle and weak performance on each trait
- Gather another sample of students' work
- Use the criteria to separate work into the levels of the rubric. Decide if the rubric "works" and if the criteria help to make accurate judgments about students' work
- Find samples of student work which are good examples of strong, weak and mid-range (benchmarks). These will be useful to share with students.
- Revise the criteria. Try again until the rubric score captures the "quality" of the work.

Example: Research Paper Rubric to Evaluate Early/Rough Draft

Elements of the Paper	Scoring Scale
Statement of hypothesis or research question	3 Clearly stated 2 Present, but unclear 1 Incomplete or missing
Analysis of Information	3 Thorough and appropriate 2 Present, but not complete 1 Not attempted
Conclusion	3 Complete and appropriate 2 Present, but not completely appropriate 1 Not articulated or not appropriate

As the complexity of the paper develops, the complexity of the rubric may develop, too. **The goals of your assignment and intent of your assessment (formative or summative) will invite different calibrations of criteria.** The rubric above lends itself to providing quick formative feedback on a project in process. The format invites brief notes to guide the student.

Other possible scales to replace the point total may be descriptors such as:

1 = "Competent" or "Completed";

2 = "Developing" or "Approaching completion";

3 = "Insufficient development" or "Not yet completed".

More elaborate rubrics guide the final draft and summative evaluation. See next page for an example.

This rubric adapted from: Paloma, C. A. and Banta, T. W. (1999). Assessment Essentials: Planning, Implementing and Improving Assessment in Higher Education. San Francisco: Jossey-Bass, p. 164

Example: Research Paper Rubric to Evaluate Final Draft

Elements of the Paper	Scoring Scale
Review of Relevant Literature	3 Well-defined, thorough, well-organized 2 Generally appropriate focus; some extraneous material or key sources missed 1 Merely lists or reports studies; little or no logic to selection of sources
Statement of Hypothesis or Research Question/s	3 Elegant and directly related to literature 2 Present, but unclear; logic of the questions is not explicit 1 Not present or poorly constructed, unclear, not related to literature
Methodology Description	3 Correct method selected; detailed description of method; logically connected to question/s 2 Weak method or insufficient description of method; suggests lack of capacity to do the study 1 Inappropriate method selected
Presentation of Results	3 Data are properly reported; well-organized; complete data set relevant to study 2 Data are appropriate but mistakes in presentation are evident; may be less than complete 1 Data are missing or improperly reported; poorly organized and misleading; incomplete
Discussion of Findings	3 Discussion is insightful, thorough, well-organized 2 Discussion is mechanical; some gaps in analysis; organization may be weak 1 Discussion fails to interpret data (merely repeats results)

Finer distinctions in some or all of the cells may be warranted. (Note: All cells do not necessarily have to have an equal number of criteria).

Resources for Rubrics

Classroom Assessment Techniques Scoring Rubrics

<http://www.flaguide.org/cat/rubrics/rubrics1.php>

Create Rubrics for your Project-Based Learning Activities

<http://rubistar.4teachers.org/index.php>

Bloom's Taxonomy of Educational Objectives (Cognitive)

Level of thinking	Skills Demonstrated
Knowledge	<ul style="list-style-type: none"> • observation and recall of information • knowledge of dates, events, places • knowledge of major ideas • mastery of subject matter • <i>Question Cues:</i> list, define, tell, describe, identify, show, label, collect, examine, tabulate, quote, name, who, when, where, etc.
Comprehension	<ul style="list-style-type: none"> • understanding information • grasp meaning • translate knowledge into new context • interpret facts, compare, contrast • order, group, infer causes • predict consequences • <i>Question Cues:</i> summarize, describe, interpret, contrast, predict, associate, distinguish, estimate, differentiate, discuss, extend
Application	<ul style="list-style-type: none"> • use information • use methods, concepts, theories in new situations • solve problems using required skills or knowledge • <i>Questions Cues:</i> apply, demonstrate, calculate, complete, illustrate, show, solve, examine, modify, relate, change, classify, experiment, discover
Analysis	<ul style="list-style-type: none"> • seeing patterns • organization of parts • recognition of hidden meanings • identification of components • <i>Question Cues:</i> analyze, separate, order, explain, connect, classify, arrange, divide, compare, select, explain, infer
Synthesis	<ul style="list-style-type: none"> • use old ideas to create new ones • generalize from given facts • relate knowledge from several areas • predict, draw conclusions • <i>Question Cues:</i> combine, integrate, modify, rearrange, substitute, plan,

	create, design, invent, what if?, compose, formulate, prepare, generalize, rewrite
Evaluation	<ul style="list-style-type: none"> • compare and discriminate between ideas • assess value of theories, presentations • make choices based on reasoned argument • verify value of evidence • recognize subjectivity • <i>Question Cues</i> <p>assess, decide, rank, grade, test, measure, recommend, convince, select, judge, explain, discriminate, support, conclude, compare, summarize</p>

<http://www.coun.uvic.ca/learn/program/hndouts/bloom.html>

Sample **Multiple-Choice Items That Measure at Various Levels**
On [Bloom's Taxonomy](#)

Knowledge Level

Which of the following are the raw materials for photosynthesis?

- a. Water, heat, sunlight
- b. Carbon dioxide, sunlight, oxygen
- c. Water, carbon dioxide, sunlight
- d. Sunlight, oxygen, carbohydrates
- e. Water, carbon dioxide, carbohydrates

[This asks students to remember, then “point to” specific information.]

Comprehension Level

If living cells similar to those found on earth were found on another planet where there was no molecular oxygen, which cell part would most likely be absent?

- a. Cell membrane
- b. Nucleus
- c. Mitochondria
- d. Ribosome
- e. Chromosomes

[Comprehension questions ask students to do some thing with some information—interpret it; predict, estimate, etc.]

Application Level

Phenylketonuria (PKU) is an autosomal recessive condition. About one in every fifty individuals is heterozygous for the gene but shows no symptoms of the disorder. If you select a symptom-free male and a symptom-free female at random, what is the probability that they could have a child afflicted with PKU?

- a. $(.02)(.02)(.25) = 0.0001 = 0.01\%$, or about 1/10,000
- b. $(.02)(.02) = 0.0004 = 0.04\%$, or about 1/2,500
- c. $(1)(50)(2) = 100\% = \text{all}$

d. $(1)(50)(0) = 0 = \text{none}$

e. $1/50=2\%$, or $2/100$

[This item provides some information or directs students to recall information such as a formula or method and use it (apply it) in some fashion.]

Analysis Level

Mitochondria are called the powerhouses of the cell because they make energy available for cellular metabolism. Which of the following observations *is most cogent* in supporting this concept of mitochondrial function?

a. ATP occurs in the mitochondria.

b. Mitochondria have a double membrane.

c. The enzymes of the Krebs cycle, and molecules required for terminal respiration, are found in mitochondria.

d. Mitochondria are found in almost all kinds of plant and animal cells.

e. Mitochondria abound in muscle tissue.

[Analysis questions go beyond application to require students to think more like an expert. Students need to make crucial differentiations or distinctions among options—in this case functions of interrelated parts of a cell.]

Evaluation Level

Disregarding the relative feasibility of the following procedures, which of these lines of research is likely to provide us with the most valid and direct evidence as to evolutionary relations among different species?

a. Analysis of the chemistry of stored food in female gametes

b. Analysis of the enzymes of the Krebs cycle

c. Observations of the form and arrangement of the endoplasmic reticulum

d. Comparison of details of the molecular structure of DNA

e. Determination of the total percent protein in the cells

[Evaluation questions assume students have developed or mastered a set of criteria for judgment. In this case, criteria for “valid” evidence and “direct” evidence regarding evolutionary relationships of species need to be employed.]

Single Concept Measured at Different Levels

(Topic: “Reliability” in testing)

Knowledge Level

The split-half technique is used to establish a test's

a. Reliability

b. Validity

c. Objectivity

d. Correlation with a criterion

Comprehension Level

If a test is doubled in length by adding comparable items, the reliability coefficient

a. Remains the same

b. Will be doubled

c. Will increase by some amount

d. May or may not increase

Application Level

An odd-even correlation coefficient of .80 is obtained for a 100-item test. What is the estimated reliability of the entire test?

- a. .64
- b. .78
- c. .89
- d. .91

Analysis Level

Which of the following steps would most likely lead to an increase in the reliability estimate for a test?

- a. Increasing the number of persons tested from 500 to 1,000
- b. Selecting items so that half were very difficult and half very easy
- c. Increasing the length of the test with more of the same kinds of items
- d. Increasing the homogeneity of the group of subjects tested

Synthesis Level

An instructor plans to use the split-half method to obtain an estimate of the reliability of a 100-item speed test covering math computation. Explain to this instructor why the split-half method should *not* be used; recommend an alternative procedure and defend your choice.

Evaluation Level

An instructor in an introductory psychology course administered a 150-item multiple-choice comprehensive final exam in a course. The items had a range of difficulty from easy to hard. Which of the following would be the *best* procedure for determining the reliability of the test?

- a. The coefficient of equivalence
- b. The coefficient of stability
- c. Kuder-Richardson formula 21
- d. Split-half length

Quick Bits on Testing

If you use exams, three or four plus a final in a semester seems optimal for student success.

Give the first test around week 3 or 4 of the semester.

On objective tests, items that are answered correctly by about 50%-70% of students are optimal in giving a wider (better) distribution of scores. (Use these percentages as a tool for fast "item analysis." A class average of <50% or >70% on an item merits review of the item.)

Start each test with a couple easier questions so as not to discourage students.

Allocate about 30 seconds per T/F item; 1 minute per multiple choice item (if measuring higher order thinking); 1 minute for fill-in blank; 2 minutes for short answer items;10-15 minutes for short essay; 30 minutes for essay of about three pages, hand written.

Jacobs and Chase do not advocate providing choices of essay questions or including extra-credit items.

When grading essay exams, write out what you think are the key terms, concepts, relationships, points, etc. that should be in the answer. Award points based on their proper appearance in the essay. [A "criterion-based" approach]

Or

Compare essays and rank them in clusters, award relative grades [A "norm-based" approach]

Student bluffing techniques on written assignments:

- Writing on all questions even when they know little or nothing
- Stressing the importance of the question
- Blatantly agreeing with the professor
- Name dropping without details
- Writing on related points but missing or avoiding the central issue/s of the question
- Writing in general terms ("Some say ...; One cannot always agree...")

True/False items can be constructed to contain a number of statements such as:

The Boston Tea Party (1773) was:

T/F Carried out by Native Americans

T/F Planned as a revolt against taxes

T/F Don because the tea market in the US was glutted

When writing T/F items, avoid specific determiners ("all", "always", etc) or indefinites ("A long time ago"; "many", etc.).

When writing T/F items, state items positively; avoid negative statements.

Students are actually good sources of questions; enlist their help in creating test banks.

When discussing tests or quizzes, you can avoid unpleasant arguments by:

- Acknowledge that problems of interpretation are inherent in human communication
- To accommodate anticipated problems of interpretation, you do an item analysis
- You may award the points for the bad items (e.g. three items that are outside the <50% to >70% range mean the equivalent number of points automatically awarded) or you delete the items or you may delete the related points from the base (e.g. deleting three items worth 2 points each moves the base score from 100 to 94) or some other system you feel is fair. (Grading is thus more objective.)
- Then discuss the items you wish to discuss as prompts for learning or reinforcing the content.

Where possible, use the test as a learning experience. For example, on a math exam, when you hand it back, you may offer students a chance to identify some specific questions and explain why they missed the question. Some compensatory points can be awarded for appropriate explanations.

Other resources

Jacobs, L. C. & Chase, C. I. (1992). *Developing and Using Tests Effectively*. San Francisco: Jossey-Bass. (CSUS Library/ 3 North, LB 2366.2 .J33 1992)

A Handbook for Improving Test Construction Skills
http://www.indiana.edu/~best/pdf_docs/better_tests.pdf

Test Construction: Some Practical ideas
<http://www.utexas.edu/academic/cte/sourcebook/tests.html>

Authentic Assessment Toolbox
<http://jonathan.mueller.faculty.noctrl.edu/toolbox/tasks.htm>

Online Assessment Resources for Teachers
<http://www.uwstout.edu/soe/profdev/assess.shtml>

Integration of the Disciplines Authentic Assessment
<http://oregonstate.edu/instruction/ed555/zone5/zone5hom.htm>

Incorporating Authentic Assessment
<http://www.park.edu/cetl/quicktips/authassess.html>