1

Classroom assessment information should be the basis for important classroom processes and outcomes: students' study and work patterns, students' understanding of what they are learning, and teachers' instructional and grading decisions. Attention to principles of assessment quality, especially validity and reliability, increases confidence in the quality of assessment information.

Assessment Theory for College Classrooms

Susan M. Brookhart

Before any evaluation methods are selected and used in a course—alternative methods like those described in this volume or conventional methods like exams and term papers—it is worth stopping to think about assessment theory. Readers of this volume should have less trouble with that assertion than almost any other audience I can think of. Academically trained people recognize that in any field, things do not "just happen." In assessment as in any other field, there are some theoretical principles that will help you organize what you do; separate good practices from bad ones; and especially, recognize, appreciate, and use the information you get from your classroom assessments.

This general introduction to assessment theory, written with the college classroom context in mind, is intended to give you some tools you can use as you apply the assessment methods in this volume and as you find or create other assessment methods. Assessment practices based in sound theory will lead to high-quality information about student achievement in the classroom.

What Is Assessment For?

Assessment, broadly defined, means collecting information about something to be used for some purpose. It is a broader term than *measurement,* which means applying a set of rules (some score scale) to an attribute of something or someone to obtain quantitative information about it (a score or number of some kind). Assessment can include measurement. For example, when you use a multiple-choice exam to measure student achievement of a set of

knowledge and skills developed in your course, you typically set up a scale with rules like one point for each right answer. The result is a score that places each student on a scale of achievement. Assessment can also include collecting qualitative information—for example, when you ask a student to describe to you what information from a text he or she found difficult and why. Both kinds of information, quantitative and qualitative, can be useful assessment information. Which to use and why depend on the purpose of your assessment and what you plan to do with the information. Typical classroom assessment purposes include providing feedback to students for their studying, making instructional decisions (what to emphasize in the next lessons and in what manner), assigning grades, and advising students about additional coursework.

Evaluation goes one step further. Evaluation means using assessment information to make judgments about the worth of something. Notice the root "valu" in the middle of the word. Examples of value judgments instructors make based on assessment information include deciding students do or do not know enough about a topic to go on to the next, deciding that a particular lecture or group project was (or was not) worth repeating next year, passing or failing a student, and recommending a student for special work.

If you give a midterm exam and a student scores 64 percent, that is both a measurement and an assessment. If you use that information to conclude that your student should come see you to get extra help or remedial assignments, that is evaluation. If you ask what the problem seems to be, the student's response is also assessment information but not measurement (no numerical scale). Your judgment about the worth of the student's insights is evaluation. Your take on how you should work on the problem together involves both evaluation and instructional decision making—and, one hopes, additional ongoing assessment.

Formative and Summative Assessment

Formative assessment gives assessment information that is useful for continued student learning, positive classroom change, and other improvements. Summative assessment gives assessment information that is useful for making final decisions: for example, assigning end-of-term grades. This sounds like a neat distinction, but in classroom use the boundaries blur, for a couple of reasons. First, formative and summative assessments describe two assessment functions. That is, they describe the *use* of assessment information. Whereas some information is more conducive to being used formatively and some is more conducive to being used summatively, it is the use and not the information that makes the distinction.

The same information can be used for both functions. For example, you might use final exam scores in assigning your course grades and also use them to make modifications to the course content or to the exam itself for the next term. Or you might use midterm exam scores as part of your

course grade, and a student might also use the information to change the way he or she studies. If I gave you a copy of a test or a description of a project or paper assignment, you would not be able to tell whether it was a formative or a summative assessment. You would only know that by asking me what I did with the information about student achievement yielded by the assessment. There is evidence that no matter what instructors intend, good students will try to use any information about their achievement in a formative way for their own future (Brookhart, 2001). That is part of what distinguishes good learners.

How Can I Assess the Quality of Students' Work?

To assess the quality of your students' work, you need to know what assessment options are available to you, how to construct or select an appropriate assessment from these options, how to get these assessments to yield good-quality information, how to interpret the information and help students to interpret it, and how to use the information yourself and help students (and sometimes others) to use it. You also need to follow this cycle through to the end so that the information does get used; otherwise, the students' time and yours are wasted.

Types of Assessments. There are four basic ways to collect assessment information (three if you count portfolios as a collection of other assessment methods): paper-and-pencil assessments, performance assessments, assessments based on oral communication, and portfolios. Three different kinds of assessment information feedback can be generated for each: objectively scored numerical data, subjectively scored numerical data, and written feedback. Three types of feedback times four types of assessments gives twelve basic categories to choose from, with a lot of variation within each one! Not to worry, though. Knowing the range of options you have to choose from actually makes deciding on an assessment easier. Once you know what content domain you are assessing and what the purpose is, choosing an assessment becomes a matter of finding the best kind of assessment for its intended use. Then designing the specific assessment is less like staring at a blank screen and more like "writing to specifications."

Assessment Type 1. Paper-and-pencil assessments include objective item tests that use multiple choice, true or false, matching, and fill-in items as well as essay tests. Paper-and-pencil tests are usually given in on-demand settings, as when students "sit for" an exam.

Assessment Type 2. Performance assessments use observation and judgment to assess either a process (how the student does something) or a product (student-created work). Common performance assessments include term papers, academic or technical projects, oral reports, and group demonstrations.

Assessment Type 3. Oral communication is an often-forgotten assessment method. Its most common use in college classrooms is for formative

assessment during instruction, when the instructor asks students questions in class.

Assessment Type 4. Portfolios are systematic collections of student work over time, often with accompanying student reflections. The work can be scored as a set; individual pieces of work in the portfolio can be scored; or the portfolio can be used as information for conferences, written feedback, or other communication between instructor and student.

Feedback Type 1. Objective scoring is the kind of one-right-answer scoring that anyone can do with an answer key. Objectively scored items are easy to grade but difficult to write well, and they require more instructor preparation time than subjective items.

Feedback Type 2. Subjective scoring is the kind of scoring that requires judgment. Despite the sometimes pejorative use of the term (as in, "that was so subjective!"), good academic judgment well applied is the heart of a discipline. Thoughtfully applying good rubrics or scoring schemes—ones that use clear descriptions of the work, not just evaluative terms like *excellent, good, fair,* or *poor*—is an effective way to judge quality of complex work (Arter and McTighe, 2001). If possible, share the criteria with students during (and as part of) instruction before the assignment is made.

Feedback Type 3. Written feedback is particularly good for formative assessment. If you describe to a student ways he or she could improve the work, you are providing important information for the student's growing concepts and skills.

Types of Grades, Scores, and Scales. Once you decide that you are going to use quantitative scales because you need numerical data, you need to figure out what kind of scales will give you the best information for your purpose. Again, knowing what your choices are will help.

Test Scores. If you are using a test, decide how many points each item should be worth. Actually, the best way to do it is vice versa: decide how many points each particular course objective should be worth, proportional to its importance or instructional emphasis, and then write the appropriate number of test items. Multipoint essays or show-the-work problems should have some sort of scoring scale, typically either rules for assigning points to attributes of the answer or a rubric (see below).

Analytic Versus Holistic Rubrics. Rubrics are scales, usually short ones, constructed to rate the quality of student work along a series of performance levels described under a criterion. When you apply several scales to the same work—for example, by applying both a rubric for content and one for style to a paper, you are using analytical rubrics. When you make overall judgments on one rubric, you are using a holistic rubric. The same criteria can be used either way, as shown in the example in Exhibit 1.1.

With analytical rubrics, each criterion is considered separately. With holistic rubrics, the criteria are considered simultaneously; to decide where to place a particular piece of student work, select the performance level that *best* describes the work.

Exhibit 1.1. Example of General Analytic and Holistic Rubrics, Using the Same Criteria, for a Question on an Essay Test

Analytic Rubrics (Three Criteria)

Thesis and organization
4 Thesis is defensible and stated explicitly; appropriate facts and concepts are used in a logical manner to support the argument
3 Thesis is defensible and stated explicitly; appropriate facts and concepts are used in a logical manner to support the argument, although support may be thin in places or logic may not be made clear
2 Thesis is not clearly stated; some attempt at support is made
1 No thesis or indefensible thesis; support is missing or illogical

Content knowledge
4 All relevant facts and concepts included; all accurate
3 All or most relevant facts and concepts included; inaccuracies are minor
2 Some relevant facts and concepts included; some inaccuracies
1 No facts and concepts included, or irrelevant facts and concepts included

Writing style and mechanics
4 Writing is clear and smooth; word choice and style are appropriate for the topic; no errors in grammar or usage
3 Writing is generally clear; word choice and style are appropriate for the topic; few errors in grammar or usage, and they do not interfere with meaning
2 Writing is not clear; style is poor; some errors in grammar and usage interfere with meaning
1 Writing is not clear; style is poor; many errors in grammar and usage

Holistic Rubric (Same Three Criteria)
4 Thesis is defensible and stated explicitly; appropriate facts and concepts are used in a logical manner to support the argument; all relevant facts and concepts included; all accurate. Writing is clear and smooth; word choice and style are appropriate for the topic; no errors in grammar or usage
3 Thesis is defensible and stated explicitly; appropriate facts and concepts are used in a logical manner to support the argument, although support may be thin in places or logic may not be made clear. All or most relevant facts and concepts included; inaccuracies are minor. Writing is generally clear; word choice and style are appropriate for the topic; few errors in grammar or usage, and they do not interfere with meaning
2 Thesis is not clearly stated; some attempt at support is made; some relevant facts and concepts included; some inaccuracies. Writing is not clear; style is poor; some errors in grammar and usage interfere with meaning
1 No thesis or indefensible thesis; support is missing or illogical; no facts and concepts included, or irrelevant facts and concepts included. Writing is not clear; style is poor; many errors in grammar and usage

Note: Numbers indicate the points assigned for each rubric.

Source: Adapted from Brookhart, 1999, pp. 47–48. Used by permission.

Analytical rubrics have the advantage of giving more information to both instructor and student. Use analytical rubrics if you want the student to be able to glean diagnostic information by seeing several scores on different attributes of the work. Analytical rubrics have the disadvantage that grading takes longer than with holistic rubrics.

Holistic rubrics have the advantage of speed because only one global judgment is required to arrive at a score. They are therefore better for grading and other summative purposes than for formative purposes. They have the disadvantage of not, by themselves, giving much information about exactly what was thorough or skimpy, clear or unclear, accurate or inaccurate, well reasoned or poorly reasoned, and so forth, about the work.

General Versus Task-Specific Rubrics. General rubrics are those that describe levels of performance for a whole set of similar performance tasks. For example, the rubrics in Exhibit 1.1 can be used with many different essay assignments. General rubrics are recommended because they can be shared with the student ahead of time, thus being part of instruction and also giving students clear information about scoring ahead of time. They take a little longer to learn and to use reliably than task-specific rubrics, but their instructional value is usually worth the trouble. In some senses, you want students to carry around in their heads the definition of general good work found in the rubric; that in itself can be part of the learning.

Task-specific rubrics have elements of the specific problem in them and thus cannot be shared with students ahead of time because they give away the desired answer. An example would be: "Students get a 4 on this problem if they correctly identify Sam as the fastest runner, with a speed of 11.76 minutes, and have one of the following correct explanations (which would be listed)." Task-specific rubrics are easy to use quickly, so scoring is speedy, but you need to write a new rubric for every problem. Use them only when the main purpose of scoring is to ensure that responses contain certain specific facts. Sometimes, instead of a task-specific rubric, it is easier to use a scoring scheme that simply awards points to various required parts of the essay or performance.

Norm- Versus Criterion-Referenced Scales. Any score scale makes an implicit comparison between the work scored and either other students' work (for example, this paper is better or worse than that paper) or some kind of performance standard (for example, this paper has a good thesis logically supported by a variety of evidence and examples). Assessments that yield scores that compare students' work with that of other students are called "norm-referenced" assessments. Assessments that yield scores that compare students' work with a standard are called "criterion-referenced" assessments. For most classroom assessment purposes, you want criterion-referenced scores that tell students how they did with the course material, not other students.

Ordinal Versus Interval Level Measures. Different kinds of scales use different levels of measurement. Rubrics and other, typically short, scales that

describe a continuum of achievement quality are ordinal level measures. Test scores and other, typically longer, ways of adding up points for work are interval level measures. This is important because you need to take into account what kind of "data" you will have so you know what you can do with it. The best way to "average" interval level measures is to use the mean. The best way to "average" ordinal level measures is to use the median. If you need to put long and short scales together—for example, for a final grade—you need to find a method that preserves the meaning of the performance information from both scales. Readers who want to learn more about putting different scales together are referred to Brookhart (1999).

Types of Non-Numeric Feedback. Every assignment students do should receive some sort of feedback, but not every assignment needs a score. Sometimes teacher-written feedback (for example, on drafts of essays or preliminary designs for projects) is the most appropriate feedback. Oral feedback can be helpful, too, but with the number of students most instructors deal with, written feedback is recommended. It is difficult to keep too many different comments straight in one's memory! Sometimes other students' responses to the work of their peers make helpful feedback. This can be done orally (for example, in paired or small-group activities in class) or in writing.

Good verbal feedback describes to the student the qualities of the work submitted and makes constructive suggestions for improvement. Good verbal feedback leaves room for student choice in the improvement. I once had a student turn in a report making only the changes I had noted. That was my fault! Instead of writing feedback, I did copyediting. The student did not learn anything further about the qualities of a good report.

How Do I Know My Assessments Give Me Good Information?

Do not make the mistake of believing that alternative assessments are all good or that conventional assessments are all bad, or vice versa. General principles of information quality apply to all assessment information, although these principles may play out a little differently for different types of assessments. For classroom assessment, the two most important indicators of assessment information quality are called validity and reliability (Brookhart, 2003). A third indicator, feasibility or utility, is important in practice, too. An assessment that will take more time than you have, for example, is not much help. Other important assessment qualities include fairness, use of appropriate score scales (discussed above), and appropriate administration and reporting (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

Various authors have recommended ways to collect evidence for the validity and reliability of classroom assessments. What is presented here is

not an exhaustive treatment but, rather, practical recommendations about the kinds of evidence an instructor can routinely collect. You will find that if you have evidence that your assessments are valid and reliable, you will feel confident about acting on the basis of their results, whether those actions are formative (for example, going back over an unclear concept) or summative (that is, grading).

Validity. Validity of assessment information refers to its meaning and value. Assessment information should mean what it is supposed to mean. Sound obvious? What about the college course in poetry where much of the term was spent interpreting poems and understanding imagery and then assessed with an exam that had a big "match the poets to their poems" section? The instructor needs a measure of achievement of interpretation and gets a measure of achievement of author-title memorization.

Relating assessment information to course objectives gives important evidence for the validity of your course assessments (Walvoord and Anderson, 1998). This may sound like common sense, but it is important to do thoroughly and thoughtfully. It is not enough to know that the topics on your exams and projects match your course objectives. It is also important to know that the depth of knowledge required and the cognitive level of the tasks (recall or higher-order thinking) match. Finally, the proportions of the various topics and thinking levels on your assessments should go together with the same emphasis you intended for your instruction.

If this all lines up, you will probably find that you have another source of evidence for the validity of your assessments: good consequences for learning and instruction (Moss, 2003). Does your exam point studying to "the right stuff" (as opposed to trivia, or points you did not intend to emphasize)? Does your paper, project, or other assignment result in sharpening the skills (research, writing, and the like) that you intended to teach? Positive intended consequences for learning and minimal negative unintended consequences can be interpreted as evidence for the validity of course assessments.

Reliability. Reliability in achievement measures refers to the amount of confidence you have that the score the student obtained is his or her actual level of achievement. Of course, no measure is perfect. A small margin of error is expected and tolerated. However, if measurement error gets too large, the score information is not useful.

In classroom assessment, there are several reliability concerns, that is, several places where measurement error can creep in. For all subjectively scored work, and for all work where written judgments are rendered, rater accuracy is a concern. Would another person look at the work and draw similar conclusions about its level of quality? Everyone has stories about "easy" and "hard" graders. If the same essay graded by Mr. Smith would yield different information if read by Ms. Jones, that is a problem.

Most of the time, you will not have the time or opportunity to double-score assignments with a colleague (the "acid test" of reliability of scoring

judgments). You can maximize the accuracy of your own scoring in two ways: by having clear criteria written out ahead of time and shared with students, if possible; and by using example papers (sometimes called exemplars or anchor papers) or projects for each level of grading.

Another reliability concern is sufficiency of information (Smith, 2003). This one can be easily overlooked, especially in a course with lots to do and where there is little time to do anything twice. If you only ask one question on a test, how do you know that the student's work, whether right or wrong, accurately indicates what he or she can do? Anyone can guess right once or goof once. If you ask several questions about the same course objective, the pattern of student work begins to show—one hopes consistently—what level of work the student can do. Try to have at least five items (or five points) on any one topic before you place too much confidence in conclusions. Rules of thumb are dangerous if applied without thinking; others would say that even more points than five are required for accurate judgments. Probably in a survey course with many topics, the "five-point" rule of thumb is better than no rule of thumb. In more advanced courses that cover fewer topics in more depth, you can do much better than that.

Incorporating High-Quality Assessments into Manageable Classroom Practice

Once you begin to think in terms of the assessment principles laid out in this chapter, it does not take any longer to do high-quality assessments than it does to do poor-quality assessments. And because the information you get from high-quality assessments is better information, in the long run you will actually construct a better course, know more about what your students understand, and be more helpful when they do not understand.

Start with the basic questions, as laid out in this chapter. For every assessment purpose, ask yourself, "What information do I need?" Once you know what you need and why, ask yourself, "What would be the best way to get this information?" Answer your question by thinking through your assessment options and select the one(s) you are going to use.

Then, for each assessment, ask the basic validity and reliability questions: "Would student performance on this assessment really indicate the particular kind of achievement I need to know about?" and "Will I have enough information about each student to be sure about my conclusions?" If the answer to either one is no, adjust before you continue.

And, finally, the usefulness question: In the best case, the assessment information will be useful to you for your purposes (instruction, grading, and so forth) *and* useful to the student as feedback for learning. Sometimes that means you have to provide several kinds of information—for example, both scores and written feedback; sometimes the same information can be used for both student and instructor needs.

The more your assessments begin to provide both you and students with valuable information, the less trauma will be involved. Rodabaugh and Kravitz (1994) did a series of simulation studies and found that a professor who is perceived as fair, especially in testing procedures, will be respected, liked, and likely to be chosen for another class. A professor who is not perceived as fair will not be as respected, liked, or chosen even if he or she gives high grades. Test this out with memories from your own past; the instructors you remember most, and best, were probably not the "easy A's" or the ones who were simply sweet-tempered or charming. They were most likely the ones in whose classes you remember learning something. That learning cannot happen, at least not in a guaranteed manner for all students, without clear, accurate information about achievement—that is to say, without sound assessment.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999.

Arter, J., and McTighe, J. *Scoring Rubrics in the Classroom.* Thousand Oaks, Calif.: Corwin Press, 2001.

Brookhart, S. M. *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy*. ASHE-ERIC Higher Education Report, Vol. 27, No. 1. Washington, D.C.: George Washington University Graduate School of Education and Human Development, 1999.

Brookhart, S. M. "Successful Students' Formative and Summative Uses of Assessment Information." *Assessment in Education,* 2001, *8*(2), 153–169.

Brookhart, S. M. "Developing Measurement Theory for Classroom Assessment Purposes and Uses." *Educational Measurement: Issues and Practice,* 2003, *22*(4), 5–12.

Moss, P. A. "Reconceptualizing Validity for Classroom Assessment." *Educational Measurement: Issues and Practice,* 2003, *22*(4), 13–25.

Rodabaugh, R. C., and Kravitz, D. A. "Effects of Procedural Fairness on Student Judgments of Professors." *Journal on Excellence in College Teaching,* 1994, *5*(2), 67–83.

Smith, J. K. "Reconsidering Reliability in Classroom Assessment." *Educational Measurement: Issues and Practice,* 2003, *22*(4), 26–33.

Walvoord, B. E., and Anderson, V. J. *Effective Grading: A Tool for Learning and Assessment.* San Francisco: Jossey-Bass, 1998.

SUSAN M. BROOKHART is an educational consultant based in Helena, Montana, and an adjunct professor at Duquesne University in Pittsburgh, Pennsylvania.