Assessing Student Learning

A Common Sense Guide

SECOND EDITION

LINDA SUSKIE

FOREWORD BY TRUDY W. BANTA

Readers should be aware that Internet Web sites offered as citations and/or sources for further information may have changed or disappeared between the time this was written and when it is read.

Jossey-Bass books and products are available through most bookstores. To contact Jossey-Bass directly call our Customer Care Department within the U.S. at 800-956-7739, outside the U.S. at 317-572-3986, or fax 317-572-4002.

Jossey-Bass also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

CHAPTER 3

What Is Good Assessment?

Some Valuable Ideas You'll Find in This Chapter

- The best assessments are those whose results are used to improve teaching and learning and inform planning and budgeting decisions.

- The greater the variety of assessment evidence, the more confidently you can make inferences about student learning.

- Students should have multiple opportunities to develop and achieve key learning goals.

- It is unfair to place full responsibility for a key program or institutional goal on one faculty member or one course.

- Assessment is a perpetual work in progress.

More than anything else, a good assessment is one whose results are used to improve teaching and learning and inform planning and budgeting decisions (Chapter Eighteen). In order for results to be used with confidence in these ways, assessments must have the four characteristics in Table 3.1.

This chapter focuses on just the first of these four criteria: designing assessments to yield reasonably accurate and truthful information on what students have learned. Chapter Four discusses assessment purpose, and Chapter Five examines ways to engage faculty and staff. Student learning goals are discussed in detail in Chapter Eight, but clear and important student learning goals are the underlying foundation of each of the other criteria

Table 3.1. The Four Characteristics of Useful Assessments

They *yield reasonably accurate and truthful information* on what students have learned, so that we can use the assessment results with confidence to make plans and decisions.

They *have a clear purpose,* so that the assessment results are valued and don't end up sitting on a shelf.

They *engage faculty and staff,* so the assessment becomes a useful part of the fabric of campus life.

They flow from and focus on *clear and important student learning goals,* so the results provide information on matters the college or university cares about.

and consequently are addressed in all of these chapters—and, indeed, throughout most of this book.

Is It Possible to Assess Completely Accurately?

No, it's not possible to determine with complete confidence exactly what students have and haven't learned, because we can't get inside their heads to find out what they truly know and what they don't. The best we can do is to look at samples of their behavior— what they write, produce, say, and perform—and from those samples try to estimate or infer what they truly know. Even under the best of circumstances, making an inference from these snapshots of behavior is bound to be at least somewhat inaccurate because of what psychometricians call *measurement error*—fluctuations in human performance that we can't completely control—such as:

- Whether a student is ill on the day she completes an assignment or takes a test
- Whether a student is preoccupied with an argument he's had and isn't focusing sufficiently to do his best
- Memory fluctuations (we all periodically blank out on key names and facts)
- Luck in whether a particular assignment or test question focuses on something a student knows well (we all learn some aspects of a subject better than others)
- Luck in guessing on multiple-choice questions
- Mental set (sometimes we have flashes of insight; sometimes we seem inexplicably in a mental rut)

A perfect assessment—one giving absolutely accurate information on what students have learned—thus does not exist. As Carol Geary Schneider and Lee Shulman (2007) have noted, "One of the most dangerous and persistent myths in American education is that the challenges of assessing student learning will be met if only the

right instrument can be found" (p. vii). We must instead simply strive to make assessments sufficiently truthful that we will have reasonable confidence in our findings and can use them with enough assurance to make decisions about goals, curricula, and teaching strategies. The approaches discussed in this chapter will help.

Start with Clear and Important Goals

Assessments yield reasonably accurate, truthful results and are used only if they truly assess what we want them to assess: our learning goals for our students. This is why it is critical to begin with clear statements of the most important things you want students to learn from your course or program. Chapter Eight discusses how to articulate clear statements of student learning goals.

Next, plan your assessments carefully to make sure they assess the important goals that you've articulated. Aim not only to assess your key learning goals but to do so in a balanced, representative way. If your goals are that students understand what happened during a particular historical period and evaluate the decisions key individuals made during that period, your test should balance questions on basic conceptual understanding with questions assessing evaluation skills.

If the assessment is a test, plan the test by creating a test blueprint: a list of the key learning goals to be assessed by the test and the number of points or questions to be devoted to each learning goal (Chapter Eleven). Then write the test questions so each clearly corresponds to the learning goal you've identified for it in your test blueprint. This creates a fair, balanced test.

Similarly, before creating an assignment, write a scoring guide or rubric (Chapter Nine): a list of the key things you want students to learn by completing the assignment and to demonstrate on the completed assignment. Then write the assignment itself, making sure that it will elicit from students what you are looking for.

Finally, have clear, appropriate standards for acceptable and exemplary student performance (Chapters Nine and Fifteen). If you are evaluating student papers, for example, have a clear sense of the characteristics of outstanding, adequate, and poor work.

Use a Variety of Assessments

Because any one assessment is imperfect and imprecise, collect more than one kind of evidence of what students have learned. The greater the variety of evidence, the more confidently you can

infer that students have indeed learned what you want them to. Lee Shulman (2007) calls this a "union of insufficiencies" (p. 24). Instead of assessing students solely through multiple-choice tests or writing assignments, assess them using a combination of tests, writing assignments, and other projects. One assignment might be a panel presentation, another a chart or diagram, and a third a written critique. Students might convey the essence of a novel's protagonist through a diagram, video, or oral presentation rather than only through a traditional essay. If you are assessing learning across an entire program, rather than give students just one culminating examination, look at samples of papers students have written or perhaps internship supervisors' ratings of their skills. Table 10.1 offers other suggestions for varying assignments.

Using a variety of assessments acknowledges the variety of prior knowledge, cultural experiences, and learning styles that students bring to the classroom:

- Maria is not a strong writer but can easily visualize concepts. She will better demonstrate her understanding of a complex concept if she can draw a diagram rather than write an explanation.

- Robert's culture values collaboration, and he learns more from working with others than by studying alone. He will better demonstrate his understanding if he can work with others on a group presentation rather than make a solo presentation.

- Janice is not a good test taker but is very creative. She will better demonstrate her understanding if she can create a video explaining a complex concept rather than take a test.

- Jason was home-schooled in a home without a computer, so he is still insecure using a computer. He will better demonstrate his understanding on a paper-and-pencil test than on a computer-based test.

- Leah attended a high school that stressed rote memorization and drill. She will better demonstrate her knowledge of key events in American history on a fill-in-the-blank test than in a term paper that requires critical thinking skills.

- Omar has poor test-taking skills. If question 2 stumps him, he'll likely spend the whole testing period pondering that question and never answer the remaining ones. He will better demonstrate his understanding by writing a term paper than by taking a multiple-choice test.

Thus, if all your course assignments are oral presentations, you may unfairly penalize those who have truly mastered the

material but are poor speakers, unless one of your major learning goals is to strengthen oral communication skills.

As you plan multiple assessment strategies, be sure to include, if possible, direct—tangible, visible, and self-explanatory—evidence of student learning such as samples of student work. As discussed in Chapter Two, this kind of evidence is more compelling than indirect evidence such as surveys or self-ratings. Because we are preparing students to lead productive and fulfilling lives, assessments that mirror real-world experiences can be especially useful.

Multiple assessment strategies are especially important for goals to instill attitudes and values because direct evidence of these goals is often difficult or impossible to collect. Chapter Twelve discusses strategies to assess these traits.

Choose and Create Fair and Unbiased Assessment Strategies___

No one wants to use an assessment tool with obvious stereotyping or offensive material. But it's easy to use tools that inadvertently favor some students over others. The following tips minimize the possibility of inequities:

Don't rush. Assessments yielding fair and reasonably accurate and truthful results take some time and thought to choose or create. The maxim "garbage in, garbage out" applies here. Assessments thrown together at the last minute invariably include flaws that greatly affect the fairness, accuracy, and usefulness of the results.

Aim for assignments and questions that are crystal clear. If students find a question difficult to understand, they may answer what they think is the spirit of the question rather than the question itself, which may not match your intent. Creating clear assignments, test questions, and survey questions is discussed further in Chapters Ten, Eleven, and Twelve, respectively.

Guard against unintended bias. A fair and unbiased assessment tool describes activities that are equally familiar to all and uses words that have common meanings to all. A test question on quantitative skills that asks students to analyze football statistics might not be fair to women, and using scenarios involving business situations may be biased against students studying the humanities, unless you are specifically assessing understanding of these topics. One way to detect some kinds of potential bias, especially in surveys and interviews, is to ask, "If someone were hoping to see the exact opposite of the results that I'm hoping for, would he or she

conduct the same assessment in the same way?" If you're trying to collect information to support the need for increased funding for library materials, for example, imagine you're trying to cut back on funding (difficult though this may be for you!). Would you still ask the same questions and phrase them the same way?

Ask a variety of people with diverse perspectives to review assessment tools before implementing them. This helps make sure the tools are clear, appear to assess what you want them to, and don't favor people of a particular background. Chapter Five discusses the importance of collaboration, and here it is especially important. If you are developing an assessment for a course, share it with those teaching the same, similar, or related courses such as the next course in a sequence. You'll help promote communication on what you're all collectively trying to accomplish. If you are considering published tests or surveys, Chapter Fourteen offers questions to consider as you review them.

Try out assessment tools with a small group of students before using them on a larger scale. Consider asking some students to "think out loud" as they answer a test question; their thought processes should match those you intended. Read students' responses to make sure their answers make sense, and ask students if they found anything unclear or confusing.

Attitudes, values, and the like are particularly difficult to assess fairly and accurately. Chapter Twelve offers specific suggestions on ways to do so.

Conduct Assessments Ethically

A number of professional organizations engaged in the assessment of human performance have developed statements of ethical standards. Virtually all of these statements agree on several principles of good practice for conducting assessments. Principles of good practice for sharing and using assessment results fairly, ethically, and responsibly are discussed in Chapters Seventeen and Eighteen, respectively.

Inform Students of the Nature, Purpose, and Results of Each Assessment

Students should be fully informed about each assessment and how it fits into their overall learning experience. Peter Ewell (1996) has noted that we often operate with four curricula—designed,

delivered, expectational, and experienced—that may not be congruent. Increasing this congruence through communication with students is thus not only an ethical practice but also good pedagogy. Students learn more effectively when they understand the goals, rationale, and structure of courses and programs (see Table 18.3).

Help students understand what and why they are learning by sharing answers to the questions in Table 8.6, along with your learning goals, rubrics, and test blueprints. It may be helpful to revise Web pages, catalogue descriptions, program brochures, course syllabi, and the like to make this information clear and easy to find.

If students have add-on graduation or program completion requirements, such as compiling a portfolio, completing a survey, participating in a focus group, or taking a comprehensive examination, inform them of this in writing and as early in their program as possible, as discussed in Chapter Two.

Assessment activities can be valuable learning opportunities for students only if they receive prompt, concrete feedback on their performance (Butler & McMunn, 2006). To require students to participate in an assessment activity and not give them feedback on their performance diminishes the overall value of the assessment experience and is inconsiderate of their contributions to an assessment effort.

Protect the Privacy and Dignity of Those Who Are Assessed

Take appropriate security precautions before, while, and after you conduct an assessment, and protect the confidentiality of individually identifiable information. Password-protect computer files with identifiable information, and store paper records with identifiable information in locked file cabinets. If several people are reviewing samples of student work or accessing a computer file, removing information that identifies individuals may be a wise precaution.

While it is important to protect student privacy, faculty and staff must have sufficient information to be able to do their jobs, and this can often mean sharing identifiable information. Some faculty and staff, for example, periodically hold department meetings to discuss the progress of each of the students in their program. They also consult with their colleagues about their students less formally; a faculty member concerned about a student's slipping performance might consult with the student's advisor for ideas on how to help this student get back on track. Faculty and staff are simply carrying out an important part of their responsibilities

when they hold such conversations, and considering identifiable assessment results can make the conversations more fruitful.

Give Students Ample Opportunities to Learn the Skills Needed for the Assessment

The second step of the assessment process (Chapter One) is ensuring that every student in your course, program, or college has sufficient opportunity to achieve every fundamental goal that you've articulated. All students, no matter what curricular and cocurricular choices they make, deserve to have confidence that if they complete their work successfully, they will be prepared for what lies next, be it the next course in a sequence, a subsequent program, or their life work. This is called *curricular alignment* and is discussed in Chapter Seven.

If you are truly serious about ensuring that students achieve key learning goals, design the curriculum to ensure that students have multiple opportunities to develop and achieve those goals. It is simply not fair to place full responsibility for student achievement of a major goal on just one assignment, one faculty member, or one required course.

Multiple learning opportunities in courses. We all learn best with practice (see Table 18.3), so give students repeated, purposeful opportunities to learn the major concepts and skills that they will be assessed and graded on. If one of your course goals is that students develop an appreciation of other cultures, for example, include in your syllabus several assignments and classwork specifically designed to help students develop this appreciation.

Multiple learning opportunities in programs. Students should have repeated opportunities to achieve major program goals throughout the program. If a program goal is that students write effectively, for example, the curriculum should ensure that all students, regardless of curricular choices, take multiple courses in which they learn how the discipline defines good writing, learn how to write in the discipline, and receive constructive feedback on their writing.

Multiple learning opportunities for institutional goals. If an institutional goal is to instill a commitment to community service, for example, curricula and degree requirements should ensure that every student, regardless of major or extracurricular involvement, has ample opportunity to develop this commitment before graduation. Simply offering service opportunities

that students may or may not participate in, at their option, will not suffice.

Use proven pedagogies. Give students plenty of practice and feedback, and find ways to engage and encourage them. Positive contact with faculty greatly influences the performance of some students.

Evaluate Student Work Fairly, Equitably, and Consistently

No matter how carefully assessments are constructed, most remain essentially subjective (Chapter Two) and thus prone to unintentional evaluation errors and biases, as discussed at the beginning of this chapter. Rubrics (Chapter Nine) can help ensure fair, consistent evaluation of student work, but they are nonetheless subject to scoring errors and biases such as those in Table 3.2.

Table 3.3 suggests strategies to minimize scoring errors and biases and achieve greater scoring consistency. Some of these strategies are easier said than done, of course, and following all these steps can be time-consuming and expensive. (Faculty and staff

Table 3.2. Examples of Scoring Errors

Leniency errors occur when faculty and staff judge student work better than most of their colleagues would judge it.

Generosity errors occur when faculty and staff tend to use only the high end of the rating scale.

Severity errors occur when faculty and staff tend to use only the low end of the rating scale.

Central tendency errors occur when faculty and staff tend to avoid both extremes of the rating scale.

Halo effect bias occurs when faculty and staff let their general impression of a student influence their scores, perhaps giving higher scores to a student who seeks extra help or lower scores to a student who is quiet in class.

Contamination effect bias occurs when faculty and staff let irrelevant student characteristics (such as handwriting or ethnic background) influence their scores.

Similar-to-me effect bias occurs when faculty and staff give higher scores to students whom they see as similar to themselves, such as students who share their research interests.

First-impression effect bias occurs when faculty and staff's early opinions distort their overall judgment. A student who presents her outstanding research in a sloppy poster display might suffer from first-impression effect bias, as might a student whose generally excellent essay opens with a poorly constructed sentence.

Contrast effect bias occurs when faculty and staff compare a student against other students instead of against established standards. Faculty might give a rating of "unacceptable" to the worst paper they read, even though the paper meets stated minimally acceptable standards.

Rater drift occurs when faculty and staff unintentionally redefine scoring criteria over time. As faculty and staff tire while scoring student work, some get grumpy and more stringent, while others skim student work more quickly and score more leniently.

Table 3.3. Strategies to Minimize Scoring Errors and Biases

Consider using a descriptive rubric (Chapter Nine)—one that describes student achievement at each of the rubric's performance levels.

Remove or obscure identifying information from student work before it is scored. This is called *blind scoring.*

Practice scoring consistently when faculty and staff are scoring student work together. First, discuss and come to agreement on the meaning of each performance level. Then score a few samples of student work, share your scores, and discuss and resolve any differences in your ratings. Once you're reasonably sure that you're all interpreting the rubric consistently, you can begin the actual scoring.

Have each sample of student work scored independently by at least two faculty or staff members. If those two disagree on any sample of student work, have that work scored by a third person to break the tie.

Rescore the first few samples when scoring many samples of student work to guard against rater drift.

Periodically schedule a refresher scoring practice session when faculty and staff are scoring large numbers of papers, in which they all compare their scores and discuss and resolve any emerging differences.

---

may expect extra compensation for spending hours or days scoring student work beyond that in the courses they teach.) To decide if these steps are worthwhile, consider the following questions.

Do we have a problem with scoring errors and bias? Look at the scores that faculty and staff are awarding. Are they reasonably consistent across faculty and staff, or are some faculty and staff more lenient or more stringent than the majority? If there may be reason for concern, do a spot check: ask a faculty or staff member or two to rescore—blind—a few student work samples to verify that there is indeed a problem with consistency.

What are the consequences of scoring errors or bias? If the scores are simply part of several pieces of information used to inform faculty and staff about teaching and learning successes and concerns, it may not be worthwhile to invest time and resources in rigorously eliminating scoring errors and bias. But if the scores are used to help make major decisions, such as whether students graduate or whether a program continues to be funded, ensuring accurate, consistent scoring becomes extremely important.

Are the scores of sufficient quality that we can use them with confidence for their intended purpose? If you don't think so, you may need to increase your investment in scoring accuracy and consistency.

Look at Enough Evidence

Obviously, the more assessment evidence you collect and consider, the greater confidence you will have in your conclusions about student learning. Faculty and staff who look at three hundred essays will have more confidence in their conclusions about student writing skills than those who look at ten essays. But more evidence means more precious time spent collecting and examining it, so an important question is, "How much evidence is enough?"

Should You Collect Evidence from Everyone or Just a Sample?

The temptation to ask only samples of students to participate in assessments of program or institutional goals is strong. Published test and surveys such as the Collegiate Learning Assessment (Council for Aid to Education, 2006) and the National Survey of Student Engagement (Indiana University, 2007) can be quite expensive, and the cost of administering them to everyone may be unaffordable. While the direct costs of locally designed surveys, focus groups, and interviews are not as high, the time needed to administer these to all students may be prohibitive.

Another reason to ask only samples of students to participate is that a representative sample can yield information that is almost as accurate as information from everyone. Consider that professional pollsters, trying to determine the opinions of millions of people, rarely survey more than a thousand people. If you've seen the results of such surveys, you may have noticed that pollsters note an error margin of about 3 percent. This means that if a pollster finds, for example, that 76 percent of the public think an elected official is doing a good job, the pollster is very sure (actually 95 percent sure) that if everyone could be surveyed, between 73 percent and 79 percent (76 percent plus and minus 3 percent) would say that the official is doing a good job.

The drawback of asking only a sample of students to participate is that, because participation is obviously not required, it can be very difficult to convince students to participate. Even if they do, they may not give the assessment their best thought and effort. Chapter Two discusses strategies to motivate students to participate in voluntary add-on assessments. If possible, a better strategy than inviting a sample of students to participate in an assessment is to require all students to participate and then choose a representative sample to examine.

How Much Evidence Should You Look At?

The sample size you choose depends on how great an error margin you're willing to tolerate. Table 3.4 lists the error margins of various sample sizes.

While professional pollsters often aim for samples of about a thousand people, with an error margin of 3 percent, unless your assessments may lead to major (read expensive) changes, a sample of no more than three or four hundred is probably sufficient. What if that many is still too large to be practical? What if faculty and staff have the time to score only, say, fifty or a hundred essays? You can use smaller sample sizes—whatever number you think is feasible—if you recognize that your error margin will be larger. If the assessment will likely lead only to minor changes, such as adjusting the curriculum of a course or two, a smaller sample may be fine. Consider looking at sequential samples. Start with a representative sample of, say, ten essays. Then look at another representative sample of ten essays to see if they add any new insight. If they do, look at a third sample of ten essays. Eventually you will look at a sample of ten essays that adds no new insight, and at that point, you may conclude that you have looked at enough essays.

What if you have a very small program or college? Obviously, you don't need to examine 300 papers if you have only 250 students in your course or program. Table 3.5 lists the sample sizes needed for a 5 percent error margin from some relatively small groups of students.

The ultimate answer to, "How much evidence is enough?" is to use your common sense. Collect enough evidence to feel reasonably confident that you have a representative sample of what your students have learned and can do. The sample should be large enough and representative enough that you can use the results with confidence to make decisions about a course or program. And take careful steps to ensure the accuracy and truthfulness of your assessment findings.

Table 3.4. Error Margins of Various Sample Sizes

Random Sample Size	Error Margin
9,604	1%
2,401	2%
1,067	3%
600	4%
384	5%
264	6%
196	7%

Table 3.5. Sample Sizes Needed from Small Groups
for 5 Percent Error Margins

Number of Students You Are Sampling From	Random Sample Size
1,000	278
500	217
350	184
200	132
100	80
50	44

What Is an Acceptable Participation Rate?

Just because you invite students to participate in a survey doesn't mean that they will all do so. In this age of survey fatigue, it is not uncommon to have a participation rate of only 25 percent—or less—of those invited. How high a participation rate should you aim for?

This question has both simple and complex answers. The simple answer is that survey experts have traditionally aimed to have 70 to 80 percent of those contacted participate in a survey or interview and have considered a 50 percent participation rate minimally adequate. The complex answer is that the quality of participants is more important than the quantity. In other words, having participants who are truly representative of the group from which you are sampling can be more important evidence of your assessment's credibility than its participation rate.

Imagine that faculty and staff at two large universities, each graduating about eight thousand students annually, want to learn about their seniors' self-perceptions of their thinking skills. Faculty and staff at Eastern State University send a survey to all eight thousand seniors, of which 400 are returned. Faculty and staff at Western State University send a survey to a random sample of six hundred seniors. They make strong efforts to convince students to complete and return the survey and, as a result, 360 are returned.

Which is the better approach? While more students completed Eastern State's survey, its response rate is only 5 percent. This makes it unlikely that the respondents represent all seniors. Some cohorts may be underrepresented (perhaps students in certain majors or students with certain experiences), which calls the value of the survey into question. Western State's survey yields a more respectable 60 percent return rate, which gives more confidence that the respondents are a good cross-section of all seniors, even though the number of returned surveys is smaller. Furthermore, Western State's approach may be more cost-effective;

it may be less expensive to survey six hundred students intensively than eight thousand scattershot. Relatively small-scale assessments with high participation rates may thus yield more credible results than larger assessments with low participation rates.

No matter what your participation rate is, collect demographic information on the participants. When you share the results, report the participation rate and describe how representative the participants are of the group you're surveying (Chapter Sixteen), so audience members can judge for themselves how credible the survey results are.

How Might You Choose a Representative Sample?

Samples can be selected in a variety of ways. Here are three ways particularly appropriate for student learning assessment.

"Simple" random samples.   These kinds of samples, in which every student has an equal chance of being selected, are a straightforward way to ensure that the sample is representative of all students. A simple random sample might be drawn by writing every student's name on a separate slip of paper, putting all the slips in a bag, shaking the bag, and drawing out as many names as you need. This can be done electronically by using software to generate a random sample. If such software isn't available (check with technical support staff), select students based on the last few digits of their student identification numbers, because the last digits are usually randomly distributed. If you have 250 students and wish to examine writing samples from 50 (20 percent) of them, for example, you could choose all students whose student identification numbers end in, say, 4 or 5 (20 percent of all possible digits 0 through 9).

Simple random samples aren't always practical. If you want to administer an in-class survey, for example, it wouldn't be feasible to choose a random sample of the entire student body, go to every class, and ask just the sampled students in each class to complete the survey while the rest of the students sit idle. If a simple random sample is not realistic, other kinds of samples are possible.

Cluster random samples.   These kinds of samples are taken by choosing a random sample of subgroups of students and then collecting information from everyone in those subgroups. You could take a random sample of first-year writing classes, for example, and then assess essays written by everyone in those classes. Or you could take a random sample of floors in the residence halls and interview everyone on those floors.

Purposeful or judgment samples. These kinds of samples are carefully but not randomly chosen so that, in your judgment, they are representative of the students you are assessing. Suppose that you want to assess essays written by students in first-year writing classes and would like to select a random sample of classes. Unfortunately, you know that while some faculty will gladly cooperate, others will decline to provide you with copies of student essays. You can still obtain a good sample of essays by choosing, from those classes with cooperating faculty, a sample of classes that meet on various days and at various times and seem, in your judgment, to represent a good cross-section of all first-year writing classes.

Or suppose that you want to assess student learning in general education science courses. While students may meet this requirement by taking any of seventeen courses, 80 percent take one of just two courses: Introductory Biology and Introductory Geology. Collecting assessment information from just these two courses will be far simpler and more cost-effective than collecting information from all seventeen courses and will give useful information on 80 percent of your students.

Keep in mind that if you must use a small or nonrandom sample or have a low participation rate, it's especially important to collect information showing that your sample is representative of students in general.

Consider Assessment a Perpetual Work in Progress_____

Good assessments are not once-and-done affairs. They are part of an ongoing, organized, and systematized effort to understand and improve teaching and learning.

When assessment is truly systematized, some assessment activity is happening every year. Assessments conducted just once every five or ten years take more time in the long run because there is a good chance that no one will remember, find the documentation for, or understand the rationale behind the last assessment. This means far more time is spent planning and designing a new assessment—in essence, reinventing the wheel. Imagine trying to balance your checking account once a year rather than every month or your students cramming for a final rather than studying over an entire term, and you can see how difficult and frustrating infrequent assessments can be compared to those conducted routinely.

"Systematized" does not necessarily mean doing exactly the same thing, semester after semester or year after year. Some good synonyms are *progressive, iterative,* and *spiraled.* As David Hollowell,

Michael Middaugh, and Elizabeth Sibolski (2006) have noted, a "planning document is not an immutable course of actions" (p. 7). As the needs of your students evolve in a rapidly changing world, so will your goals, curricula, teaching methods, and assessment practices. There is no point in repeating assessments that have become outdated or no longer provide new insight, so periodically take time to sit back with your colleagues and evaluate your assessment efforts and their outcomes. Chapter Nineteen discusses this further.

Should You Document Evidence of the Quality of Your Assessment Methods?

The answer depends on how the results will be used. An assessment used to make minor curricular modifications does not need thorough evidence of its quality. But assessments that help determine placement of incoming students, whether expensive modifications should be implemented, or whether a program should be terminated need more compelling evidence of their quality. Assessments whose findings are likely to be challenged (Chapter Seventeen) also need evidence of their quality.

Obviously the more rigorous and extensive your assessment evidence is, the more compelling it is, but also the more time-consuming it is to collect and evaluate. So balance the need for quality with the need for cost-effectiveness.

One important way to document the quality of assessments is to keep records of everything that has been done to maximize assessment quality. This can include reviews of assessment tools by others, tryouts of assessment strategies, rubrics used to score student work, blind scorings by colleagues, and the other strategies discussed in this chapter. Should you decide to take further steps to evaluate the quality of your assessment activities, Chapter Sixteen discusses how to do so.

The Role of Institutional Review Boards in Assessment

Title 45, Part 46 of the Code of Federal Regulations (U.S. Department of Health and Human Services, 2005) describes federal policy for the protection of human research subjects. The regulations stipulate that colleges must establish institutional review boards (IRBs) to ensure that research protects and poses no significant risk or threat to the rights and welfare of human subjects. There are three levels of review: full review (which requires appearing before the entire IRB),

expedited review (in which at least one member of the IRB reviews the research plan), and exempted from review (under which the research plan must still be sent to the IRB).

The regulations define research as "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge." Two kinds of research activities are exempt from the policy. One is research conducted in educational settings involving normal educational practices, such as research on instructional strategies or the effectiveness of teaching methods. The other exemption is for research involving the use of educational tests, surveys, interviews, and observations. The only exception to this second exemption is when the information obtained can be linked to the subjects *and* any disclosure of their responses places them at risk of liability or might damage their financial standing, employability, or reputation.

Should an assessment effort come under IRB review? Some colleges take the position that their assessment activities are "action research" (Chapter One), designed only to inform local teaching and learning practices and not to develop or contribute to generalizable knowledge. Under this position, faculty and staff are not engaged in research as defined by this policy, the assessment activities are exempt from the policy, and there is no reason to involve the IRB.

Some colleges take the position that they conduct assessments in established educational settings involving normal educational practices, or the assessments involve educational tests, surveys, and interviews that do not place subjects at risk. Under this position, assessment programs are exempt from this policy and there is no reason to involve the IRB.

Some other colleges require that assessment plans be submitted to the IRB with a formal request for exemption. Still other colleges take the position that assessment activities should undergo review. The rationale is that even though most assessments are not designed to contribute to generalizable knowledge, faculty and staff may decide later to share what they have learned with professional colleagues in a conference or journal.

When an assessment plan undergoes IRB review, the IRB requires the activities to meet the following criteria:

- Risks to subjects are minimized.
- Risks to subjects are reasonable.
- Selection of subjects is equitable.
- Informed consent is sought and documented.
- Adequate provisions are in place to ensure subjects' safety, protect their privacy, and maintain confidentiality.

Some colleges that require IRB review of assessment activities take the position that assessment activities pose minimal risk to subjects—no more than that "ordinarily encountered in daily life"—and therefore qualify for expedited review.

Obviously interpretations of the federal regulations vary, and there is no clear consensus on the role of IRBs in assessment programs. Ask the chair of your IRB for your college's interpretation.

Time to Think, Discuss, and Practice

1. The history faculty are assessing students' writing skills by evaluating senior theses for organization, focus, style, and mechanics.

   - Brainstorm three ways that the faculty might help ensure that this assessment will give them accurate, truthful information.

   - What might the faculty do to protect the privacy of the students and their professors as they conduct this assessment?

2. One of the goals of Mackenzie College's general education curriculum is for students to develop a tolerance for perspectives other than their own. Brainstorm three survey questions faculty might ask that you think would yield unbiased results.

3. The business program at Calvert College requires every student to compile a portfolio of his or her work. The program's ten faculty would like to assess student learning by examining a sample of portfolios from its two hundred graduating students. It takes about twenty-five minutes to review each portfolio. How many portfolios would you recommend that the faculty examine? Why?

Recommended Readings

The following readings are recommended along with the references cited in this chapter.

American Association for Higher Education. (1992). *Nine principles of good practice for assessing student learning*. Sterling, VA: Stylus.

Association of American Colleges and Universities. (2008). *Project on accreditation and assessment: Criteria for recognizing "good practice" in assessing liberal education as collaborative and integrative*. Washington, DC: Author.

54   Assessing Student Learning

Badger, E. (1999). Finding one's voice: A model for more equitable assessment. In A. L. Nettles & M. T. Nettles (Eds.), *Measuring up: Challenges minorities face in educational assessment* (pp. 53–69). Norwell, MA: Kluwer.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs* (2nd ed.). Thousand Oaks, CA: Sage.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education.* Washington, DC: Author.

Krejcie, R., & Morgan, D. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30*(3), 607–610.

National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement.* Madison, WI: Author.

National Research Council. (1993). *Leadership statement of nine principles on equity in educational testing and assessment.* Washington, DC: Author.

Parkes, J. (2000). The relationship between the reliability and cost of performance assessments. *Education Policy Analysis Archives, 8*(16). Retrieved September 16, 2008, from http://epaa.asu.edu/epaa/v8n16/.

Scheaffer, R. L., Mendenhall, W., & Ott, L. (1990). *Elementary survey sampling* (4th ed.). Boston: PWS-Kent.

Shavelson, R. J., & Huang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change, 35*(1), 10–19.