

Strengths and Weaknesses:

Tests, Portfolios, Interviews, Surveys, and Inventories

**by Barbara D. Wright, Assessment Coordinator, Eastern Connecticut State University
(with contributions by Craig N. Shealy, Associate Professor of Psychology and Director,
Combined-Integrated Doctoral Programs, James Madison University)**

Tests

Testing is a virtually universal part of formal education at all levels and our most familiar tool for ascertaining whether students have acquired specific cognitive knowledge or skills. Thus it is understandable that many academics, at least in the early days of the assessment movement, equated assessment of student learning with testing, particularly large-scale testing using standardized, commercially available instruments. Today, however, the limitations as well as strengths of traditional testing are widely acknowledged, and alternatives to the classic, norm-referenced, objective test are well-developed.

It is useful to distinguish between local tests constructed by an instructor or a group of faculty on a campus and commercial tests. Both commercially available and locally developed tests can be used to examine knowledge and skills acquired through a particular course, set of courses, or other experience. Beyond that, however, they offer different mixes of strengths and weaknesses.

Locally developed tests

Locally developed tests have significant advantages over commercially available ones. The most obvious is that a local test can be constructed for maximum content validity; that is, it aligns closely with the content that instructors or programs have actually covered and that they want their students to command. If the curriculum changes, the test can be quickly adapted to reflect the new content. Because of the direct connection to the campus's courses and programs, the results of local testing can provide highly relevant, useful information on students' strengths and weaknesses and thus guide improvement efforts. At the same time, a local test forestalls comparison with similar programs at other institutions—a source of anxiety for many faculty and administrators.

Most often, testing occurs in individual courses, and the level of aggregation is the individual student. However, testing can also be designed to examine cumulative content knowledge across a sequence of courses or at the culmination of an entire program of study. The results of such testing can be analyzed at higher levels of aggregation not only for what they say about individual students but also for what they show about the strengths and weaknesses of a course sequence or the program as a whole.

Designing a local test, particularly one that cuts across multiple course boundaries, requires faculty to share their course goals, pedagogy, and vision of the program—something which in practice happens too rarely in the academy. Such discussions can be enormously enlightening and are likely to lead to a more coherent program; that coherence, in turn, benefits students and strengthens the program.

The test that faculty create does not have to be in the format of the classic standardized test, i.e., a series of questions for which students select a correct answer from a set of responses that have been provided for them in multiple-choice, true-false, or matching format. Faculty are free to introduce open-ended questions, integrative exercises, or other innovations, such as space for students to write comments on multiple-choice questions or elaborate on their answers. Finally, because the local test is clearly tied to subject matter the student has studied, because it can be embedded in courses students are taking, and because it can count toward students' grades or fulfillment of graduation requirements, problems of quality of student effort do not arise.

There are also disadvantages to creation of local tests. Most importantly, test construction is not a simple task, and the likelihood that a homegrown test will contain ambiguous, poorly worded items is high. Related to this are reliability and validity issues. While content validity may be high, construct validity may be questionable and reliability uncertain. Engaging a testing company to create the test may solve technical problems, but contracting the work out is expensive, the test development process is time-

consuming, the company may insist on a multi-year contract to score and analyze results, and year-to-year modifications become complicated. Nor will a local test provide any norms for comparison with student performance at other institutions.

It is widely acknowledged that students value and study what they know will be tested; it is equally true that *how* students are tested is as important as *what* is tested. Objective, short-answer tests, whether local or purchased, run the risk of focusing on surface learning—i.e., swift recall of memorized facts and formulae—rather than on deep understanding of principles and relationships. At the same time, the short-answer format sends students the message that “education” is about recognizing “right” answers, not about actively formulating one’s own, more nuanced response. For faculty, finally, local testing may lead to singling out particular courses or instructors for blame. As a result, faculty may see assessment as a personal threat, and one of the most important process benefits of assessment—communication and enhanced collegiality—is undermined.

If testing of cognitive knowledge using a local instrument is a priority, there are ways to minimize the disadvantages and maximize the benefits. To avoid problems in test construction, faculty should make use of local expertise. Any test should be piloted before large-scale administration, and test-takers should be encouraged to comment on items that they find ambiguous or unfair. If norms are important, the local test can be supplemented with a commercial test; if the goal is to get at students’ deeper understanding, the test can employ more creative, open-ended formats, or be supplemented with other kinds of evidence. Finally, it is important to remember that assessment is ultimately not psychometric science but human judgment, and that the point ultimately is not high scores (i.e., “quality assurance”) but rather insight into what students know, or don’t, and how their learning can be strengthened (“quality improvement”). In the common search for that kind of insight, there is no place for finger-pointing.

Commercially available tests

Commercial tests have some obvious advantages. They are a traditional, widely used, and accepted means of assessment that provides external control, they are designed to be valid for specific outcomes, their items and formats have been refined over many years, and their technical quality is high. They offer norm-referenced scores, the testing company can also provide longitudinal data, and the tests’ content often reflects recent trends in the discipline or professional field. Such a test can thus provide the impetus for a program to update its curriculum. The experience of taking commercial, standardized tests also helps prepare students for licensure examinations and other high-stakes testing in professional fields after graduation. Compared with local tests, commercial tests require little time or labor from campus faculty beyond reviewing instruments and choosing the most suitable one.

Commercial tests also have distinct disadvantages, however. Precisely because faculty are not deeply involved, it is easy for them to feel disengaged from the testing, fault the results, and fail to act on them. This tendency may be reinforced by a real lack of content validity; test makers’ decisions about content may seem arbitrary even to experts in the field, and departments or programs may have very good reasons for their own unique curricular emphases. Commercial tests are less likely than local methods to stimulate productive discussion, and more likely to elicit finger-pointing, blame, and resistance. Students often see little or no connection to their studies or grades, and poor quality of effort affects their performance on the exam, if they appear for it at all. Incentives, positive and negative alike, have had mixed results.

Construct validity is another crucial concern. The short-answer, closed-response format of traditional standardized tests has been criticized as highly artificial and impoverished: it is unlike the complex, integrated tasks a graduate would normally be expected to perform once out in professional or civic life, a good score may result as much from test-taking savvy as academic knowledge or skills, and the answer selected reveals little about the problem-solving ability, critical thinking, or other competencies that lie behind that choice of answer. The scores thus have limited usefulness for improvement. At the same time, this format gives students no opportunity to construct their own answers verbally, numerically, graphically, or in other ways; nor do students have an opportunity to demonstrate important affective traits such as persistence, meticulousness, creativity, open-mindedness, or ability to tolerate ambiguity.

Testing with this kind of instrument can also reinforce unfortunate stereotypes. Many academics, for example, assume that assessment equals testing for external accountability rather than internal improvement. The format reinforces faculty bias toward an outdated “empty vessel” theory of education, in which students come to college in order to be filled up with knowledge by the experts. And finally, it reinforces students’ view of education as a process of memorizing and reproducing “right” answers, rather than learning to think independently, generate their own solutions, and distinguish among “good” and “better” answers.

The norm-referenced scores provided by test makers have long been accepted as indicators of quality at all levels of U.S. education. The problem is that such scores are entirely relative and say nothing about absolute levels of quality or achievement. One student’s or institution’s gain comes at the cost of another’s. More recently, assessment practitioners have begun to move away from norm referencing and toward clearly defined standards and criteria for knowledge and skills, with the goal of helping as many students as possible reach the highest levels of proficiency.

In addition to their dubious worth as indicators of deep complex learning, the scores from commercial tests also carry the risk of misuse and invite comparisons, however inappropriate, across institutions. Finally, such tests can also be very expensive and are unlikely to provide good value—i.e., useful information for the cost involved.

In summary, commercial tests can generate information for educational improvement, but their limitations need to be recognized and dealt with. Such tests should be supplemented with other assessment methods that reveal deeper and more complex dimensions of students’ learning and do so with greater transparency. Programs purchasing such tests should negotiate with the test maker to obtain criterion-referenced as well as norm-referenced scores and other more specific information, e.g., analysis of results for subgroups within the tested population. Finally, test results should be distributed and used with caution to avoid the anxiety and defensiveness that can doom improvement efforts.

Portfolios

Portfolios are collections of student work. They come to education from the world of artists and graphic designers, who regularly compile collections of their best work in a variety of media to show to potential employers or patrons. In other words, the artist’s portfolio is a cross-sectional view of the artist’s highest skill level at a given moment in time; it does not show longitudinal development or the creative struggle that lies behind a particular piece of work.

In the academy, portfolios have been adapted to serve simultaneously as vehicles for learning and as demonstrations of learning. They first became popular with writing programs, and that is where campus expertise in portfolios is most likely to be found. Students typically collect successive drafts to show the genesis of pieces of writing; as a culmination, students write a reflective essay in which they review their development as a writer. In other words, the typical academic portfolio is longitudinal and developmental; the reflection is often regarded by both faculty and students as the most beneficial part of the process from an educational perspective. More recently, portfolios have been introduced into a wide range of programs, from first-year experience and general education to majors and professional fields.

While the stereotypical view of a portfolio is that it is organized at the student level—i.e., an individual student collects and comments on examples of his/her individual work—the portfolio does not have to be organized in this way, as more recent examples show. There are now course-level portfolios (i.e., collections of the range of work produced in a course), program-level portfolios (collections of work that illustrate the overarching goals of a major or professional degree or general education, for example), and institutional portfolios (collections of information and examples from a whole range of the institution’s programs and resources). A wide spectrum of institutional types, from large research universities and liberal arts colleges to polytechnics and community colleges, has found the portfolio adaptable to its purposes.

In other words, portfolios have proven to be an enormously appealing and robust way to demonstrate educational quality and improvement. They can be readily adapted to different levels of analysis,

purposes, and kinds of materials. Indeed, virtually all of the other direct and even indirect assessment methods (e.g., capstone projects, performances, testing, surveys, student artifacts of all kinds) provide evidence that can be collected in portfolios and then analyzed for what they demonstrate about students' achievement of learning goals.

In contrast to most other methods, portfolios have the ability to show very graphically not just where students are but also where they came from, and how they arrived at their current level of knowledge and skill. This is essential for improvement. Portfolios clearly emphasize qualitative meaning-making over statistical analysis and privilege human judgment over a scientific approach or technical facility. Thus they are well-suited to getting at the "ineffables" particularly valued by humanistic disciplines. They are engaging and educational for both students and faculty while reducing fears of misuse.

Portfolios are not without drawbacks, however. They can be extremely labor- and time-intensive, both to compile and to review. They can be cumbersome to store, too, and quickly become unmanageable, particularly if guidelines are hazy and excessive numbers of documents are included. In order for the portfolio to work as an assessment device and not merely as a student scrapbook, criteria for learning outcomes need to be carefully defined and then applied to the review of student work. For the review to succeed, reviewers must be carefully trained in the use of criteria and rubrics, and acceptable inter-rater reliability must be established. If an electronic portfolio is used, the software can be costly, and both faculty and students require technical support and training.

Ultimately, as a qualitative approach to assessment, portfolios often must be translated into quantitative terms (i.e., through numeric scores or ratings based upon "expert" judgment), particularly if the portfolio is designed to assess specific and relevant learning processes or outcomes, which must also be articulated in measurable terms (e.g., as criteria). Likewise, as with all assessment strategies, issues of validity (e.g., does the portfolio measure what it purports to measure), and reliability (e.g., would the "quality" of a portfolio be rated similarly by different raters) must be sufficiently demonstrable, particularly if the portfolio is used for purposes of assessment research, performance appraisal, or quality enhancement at an individual, group, or institutional level.

To keep portfolios manageable, whatever the level of aggregation, samples of work can be gathered rather than collecting everything from everybody. Electronic and web-based portfolios have made storage, coding, and retrieval of selected items much easier, and there are now a number of software programs and many institutional examples from which to draw. The development of good, workable criteria and rubrics and training of readers does require a considerable ongoing investment of time and energy, but when relevant as an assessment strategy, such efforts can pay off in clearer campus communication and better learning for faculty and students alike.

Interviews

Traditional interview

Interviews are comprehensive and adaptable and can be designed to address a very wide range of outcomes. Interviews can range from highly structured activities with predetermined questions and response categories to open-ended, in-depth conversations with minimal steering from the interviewer. While structured interviews will yield quantitative data, open-ended interviews require a more qualitative, descriptive approach. What qualitative analyses lack in statistical rigor they can make up for in telling details that can provide insight and lead to improvement.

The major disadvantage of interviews is that they provide indirect evidence of learning. Students report on their satisfaction with an educational experience and their perceptions of what they have learned, how their skills have developed, or how their values have changed; but through an interview they cannot provide direct evidence of what they know or can do. Interviews can also be challenging to administer. Since useful results depend on the interviewer's expertise, training is required. In addition, students must be contacted, they must agree to participate and appear for the interview, and finally the interview itself may take considerable time. If a qualitative approach is taken, the conversation must then be transcribed for analysis.

While interviews do not provide direct evidence of what students know or can do, they can be a useful supplement to direct evidence. First, they can be constructed to help pinpoint areas where students feel dissatisfied or ill-prepared, so that assessment efforts and the use of direct methods are targeted efficiently to areas that require attention. Second, interviews and other indirect methods can reveal what lies, in Pat Hutchings' phrase, "behind outcomes," and to ask questions about what students believe helped or hindered them in their learning. The answers to those questions, in turn, can inform decisions about the kinds of changes in curriculum, pedagogy, or other aspects of the educational experience that are needed in order to improve learning. Interviews, particularly open-ended, in-depth ones, also remain one of the most practical ways to assess changes in the elusive area of values and dispositions.

Oral proficiency interview (OPI)

The Oral Proficiency Interview (OPI) was developed by the American Council on the Teaching of Foreign Languages (ACTFL) beginning in the mid-1980s; since then, the interview process, definitions of proficiency levels, and interviewer training have been refined, as have the proficiency criteria applied to reading, writing, and listening as well as speaking skills. Adaptations of the OPI have also been created which allow the interviewee to respond to video- or audio-taped stimuli.

In contrast to the traditional interview, the OPI is a direct assessment method that requires a speaker to demonstrate proficiency by responding to increasingly challenging linguistic cues. The speaker's performance is rated against an elaborate set of descriptors and classified as "novice," "intermediate," "advanced," or "superior." The descriptors or criteria for linguistic production focus on function and social context; they include such traditional emphases of language instruction as grammar, vocabulary, or pronunciation, but place them in the larger context of how well the speaker is able to communicate and behave linguistically in culturally appropriate ways.

The OPI and the criteria have had a strong positive influence on language instruction, helping make it more authentic and focused on the real point, effective communication. As an assessment tool for most academic purposes, however, the OPI is a fairly blunt instrument, one that may not register the nuances of improvement that occur over relatively short time periods for language development such as a semester or even an academic year. In addition, the OPI is highly labor-intensive, requiring 20-40 minutes per interview. Becoming a certified interviewer entails extensive training and follow-up practice as well as periodic refresher courses. For official results, interviewers are hired from outside the institution, and they charge significant fees.

More recently, ACTFL has developed *Standards for Foreign Language Learning in the Twenty-First Century* (1996, 1999). The *Standards* are organized around the "five C's"; these include "communication" but also "cultures," "connections," "comparisons," and "communities." In other words, the *Standards* offer a more robust view of language learning than the OPI alone. The 1999 edition of the *Standards* includes outcomes and suggests performance indicators for all five C's and for all levels of language learners; these can readily be used to create classroom activities and rubrics. The *Standards* thus provide a valuable alternative or supplement to the OPI for assessment of student learning in foreign languages.

Surveys

The survey is a familiar and widely accepted way of collecting information; in fact, for many institutions, assessment has been virtually synonymous with surveying and reporting results. Postsecondary institutions have traditionally made heavy use of surveys to establish student satisfaction with various aspects of their educational experience, to learn how students have fared after graduation, and to gather information from stakeholders such as employers of graduates. They may be administered in classrooms, sent by mail, or even presented to students as they assemble in cap and gown for graduation. Most often, the survey takes the form of a series of questions that are presented in written or oral form (in person or more often by telephone), but web-based surveys are growing in popularity. Most institutions have considerable survey data on hand that have never been fully analyzed, communicated, or utilized for improvement.

Surveys are almost infinitely adaptable and can be constructed to gather information on virtually any topic. Local instruments can focus sharply on local concerns, while nationally administered instruments such as the College Student Experiences Questionnaire (CSEQ) or the National Survey of Student Engagement (NSSE) can provide national and cohort norms. Surveys are relatively easy to score, if they have been carefully constructed to begin with, and most campuses have an expert on campus with experience in surveying who can provide guidance. Given the current level of assessment techniques, surveys are one of the best ways to examine students' values and attitudes.

The primary disadvantage of using surveys to assess student learning is that they provide only indirect evidence. They are self-reports from students about what they believe they have learned and what they perceive their academic skill levels to be; typically they measure satisfaction and impressions but do not generate information about deeper learning. It is good to know that a student thinks, for example, that she has acquired good writing skills in the course of her studies, but that is not the same quality of information that examining an actual example of her writing would provide. Research on the accuracy of student self-reports is inconclusive; semantic problems can arise, e.g., one student's understanding of "well-prepared" may differ substantially from another student's; a highly structured questionnaire generates information only on predetermined topics; forced choices may frustrate students whose responses do not fit any of the categories; there is a tendency for respondents to give socially approved answers; and data analysis, like good construction of the survey, is dependent on local expertise. Finally, obtaining the desired samples and an adequate response rate are perennial problems.

Surveys, like other indirect methods, can support improvement of student learning if they are carefully constructed to delve deeply into students' educational experience. They can ask, for example, in what areas students feel most or least proficient and why, and what has most helped or hindered their learning. Terms can be defined, national surveys can be customized to include open-ended questions and questions of local concern, and many institutions have developed strategies to improve the response rate and quality of response. Ultimately, however, surveys should supplement direct evidence, not replace it.

Inventories

Inventories are instruments closely related to surveys that seek to establish the presence or absence in the respondent of particular behaviors, perceptions, attitudes, or personal characteristics. Based on the response pattern, correlations may be found with past academic success, for example, or predictions may be made about future behavior, e.g., success in adjusting to life in a different linguistic community. Depending on length and the nature of the questions, the inventory may be easy or difficult to administer and score. It can be designed to be valid for specific outcomes, and the predictive information it generates can be used to coach individuals or improve programs more generally. Thousands of inventories have been developed for purposes of assessing a vast range of processes and outcomes, but not all inventories are equal in terms of quality, relevance, and accessibility. These standards are among the most important to consider when evaluating whether an existing inventory is appropriate for a particular assessment project.

By "quality," the key questions to consider are whether data regarding the reliability and validity of the inventory are available and acceptable. In the absence of such data, it is very difficult to ascertain whether and to what degree an inventory may measure what it purports to measure or do so in a manner that is predictable and consistent. By "relevance," the key questions to consider are whether the inventory can be theoretically and/or empirically linked to the goals and outcomes of the assessment project. For example, an inventory that purports to measure "empathy" may have very good reliability and validity estimates, but may be a poor predictor of whether a student will do well in his or her mathematics courses in college. However, this same inventory may be good at predicting success in working with children who are placed in out-of-home care. By "accessibility," the key questions to consider are how much an inventory will cost to use as well as how easy or difficult administration, scoring, and interpretation may be. An otherwise high-quality and relevant inventory may not be usable in the end if gaining access to it is cost prohibitive and/or it is too difficult to administer.

If no inventory can be identified for a particular assessment project—or if available inventories do not meet basic standards of quality, relevance, and accessibility—it may be possible to develop an inventory

that is specifically targeted to the assessment project. However, the process through which a high-quality, relevant, and accessible inventory is developed can be extremely time-, labor-, and resource-intensive (i.e., depending upon its nature and scope, an inventory may require access to a very large number of participants to complete the inventory during its development as well as substantial statistical and psychometric expertise over a period of several years). Alternatively, if an existing inventory can be identified that approximates these standards, it may be worth determining whether the author/owner of the inventory would consider modifications and/or further development of the inventory in order to meet project objectives in a satisfactory manner.

It is also worth noting that a quantitative measure such as an inventory or standardized test may be paired with a qualitative measure, such as an interview or portfolio to access the process or outcomes of a particular project. For example, if an assessor wanted to know how a group of students “felt” about a particular educational experience, he or she might interview these individuals upon completion of the experience and/or ask them to hand in written essays reflective of their academic work. Although very relevant information could be gathered from this approach, it might be very difficult to know whether any results or differences among these students were due to the educational experience they had *and/or* to other variables that might have interacted with the experience to influence the interview reports or essays that were produced. Among other possibilities, these interacting variables could be historical and contextual, e.g., the life experiences of the individuals or their cultural or economic backgrounds, as well as “psychological” in the broadest sense, e.g., relating to cognitive or emotional functioning. Without somehow evaluating these complex processes simultaneously, it would be difficult to know whether, to what degree, and under what circumstances the reported or observed “learning” was due to students’ educational experience and/or other variables. Along with sufficient background/demographic information, a well-developed qualitative measure that is linked to an appropriate quantitative measure such as an inventory can be ideally suited to help tease apart these complex and interacting processes, thereby resulting in much more robust, reliable, and ecologically valid assessment findings.