

Introduction: Assessing Student Achievement in General Education

Trudy W. Banta

Use of Standardized Tests

For those who followed the work of Secretary of Education Margaret Spellings' Commission on the Future of Higher Education, which issued its report in 2006, the first selection in this issue of **Assessment Update Collections** will bring a profound sense of déjà vu. From the outset, commission members considered recommending a national test—or at least state-level testing—for college students focused on communication and analytical reasoning skills. In the first selection here, Sal Corrallo reports that a similar recommendation was considered, and subsequently rejected, in the early 1990s.

The context for considering a national test for college students was actually much fuller in 1990 than in 2006, and the approach to designing the test was more careful and deliberate, involving in each of several stages academics, who ultimately would have to implement the process with students. The stage for a national test was set in 1989, when the National Governors' Association identified six goals for education. The sixth goal stated that by the year 2000 the communication, problem solving, and critical thinking skills of America's college students would increase. The governors' goals were enacted into law in 1993 as Goals 2000.

Of course the only way to tell if skills are increasing is to measure them in some way. So staff at the National Center for Education Statistics (NCES) began in 1991 to explore ways to assess the skills named by the governors. Measurement specialists and other academics with

informed perspectives on the assessment of students' learning in college were invited to write papers on the topic of national testing, and these documents were reviewed by others with similar expertise. Authors and reviewers were invited to workshops where the relevant issues were discussed and NCES staff drew conclusions about what should be done next.

NCES staff correctly concluded that a national test would have to be predicated on broad consensus regarding the definitions of communication, problem solving, and critical thinking and appropriate levels of competence for each. Accordingly, grants were awarded to qualified academics to conduct studies aimed at developing the definitions. While these studies yielded good information about the definitions that should underlie a national test for college students, Congressional representatives elected in 1994 concluded that constructing such a test would be too expensive. When Corrallo left NCES in 1996, the six-year effort to find a feasible way to assess college student learning that would yield national averages and permit comparisons across institutions came to an end.

Just as the work of Corrallo and colleagues at NCES between 1990 and 1996 undoubtedly was reviewed by members of the Spellings Commission, Corrallo and associates were influenced to some extent by the findings of research on state-wide testing of college students conducted in states like Tennessee and Washington between 1986 and 1990. In the second selection Robert M. Thorndike describes the studies carried out in the state of Washington in the mid- to late 1980s.

Washington's state Higher Education Coordinating (HEC) board proposed to require statewide testing at the end of the sophomore year. When academic administrators across the state objected to the imposition of a single test on a diverse array of institutions, the HEC board agreed to permit some pilot studies to determine the feasibility of implementing such a mandate. Three tests that were readily available were tried out in community colleges as well as baccalaureate institutions. The studies demonstrated that neither ETS's Academic Profile nor ACT's College Outcome Measures Program (COMP) or Collegiate Assessment of Academic Proficiency (CAAP) was appropriate for the purposes the HEC board staff had in mind. Faculty reviewing the tests did not find

them to be valid measures of students' communication, computation, or critical thinking skills. Moreover, the tests were not perceived to be capable of generating information that would be useful in assessing individual student progress or in evaluating curricular effectiveness.

Thorndike's report reveals other objections to the use of standardized tests of generic skills for college students—objections confirmed in studies conducted in other states in the late 1980s and early 1990s. The three standardized tests deployed in Washington were found to assess primarily reading comprehension and, to a lesser extent, computational skills. Moreover, students' scores on the tests revealed high correlations with scores on tests of entering ability (e.g., ACT and SAT scores). Administering the tests at the sophomore level had been assumed to be appropriate since prevailing wisdom held that general education requirements should have been completed by that time. As confirmed in studies conducted in other states, analysis of course-taking patterns in Washington revealed that in fact few students had completed their general education by the end of the second year in college. More damaging still, according to Thorndike, "test scores were found to be negatively correlated with the proportion of general education requirements a student had completed."

Other concerns about standardized testing of generic skills emerged from the Washington studies. In the third selection I have summarized some of these, based on my experience in shepherding the University of Tennessee, Knoxville, response to the state's performance funding initiative. Institutional missions must be considered when comparing test scores across institutions, particularly given that scores on tests of generic skills are highly correlated with measures of entering ability. Valid measures of student achievement must test what faculty in a given context are teaching. Students' motivation to do their best work must be assured if testing of any kind is to be valid. And if students are to take tests seriously, faculty must be convinced of their worth as well.

Faculty on many campuses have taken the time to analyze the content of standardized tests like the CAAP and the College Basic Academic Subjects Exam (CBASE) developed at the University of Missouri and have decided to use them to assess the components of their own

general education programs that they find to be covered by the tests. They supplement the standardized tests with other measures to provide more complete coverage of all their primary outcomes. When standardized tests are employed in this fashion—by faculty who have chosen them for their own purposes—faculty engagement in studying and applying the findings in their own classes and departments is ensured and student motivation to do well is much more likely. It is the use of standardized tests to make inferences about educational quality and institutional accountability among institutions with different missions that the studies by Thorndike and others have demonstrated to be problematic.

Use of Locally Designed Instruments

Convinced that standardized tests are not sufficient measures of all that faculty expect college students to learn, faculty across the country have developed their own assessment instruments. In her brief article, Anne G. Scott describes some early steps in this process. First faculty identify the generic skills, such as communication, technical literacy, and global competence, that all students should develop. Then they list specific outcomes associated with each of the skills. By constructing a matrix with courses on one axis and the specific outcomes on the other, faculty can check the outcomes they teach in their courses. A glance at such a matrix will reveal which outcomes are not taught at all and which are taught in so few places that students will not be likely to have sufficient opportunities to practice them. This first set of steps in developing a local approach to assessment will yield information faculty can use immediately to modify curriculum, courses, and instruction without even gathering any information from students.

Assessing Individual Generic Skills

Effective writing is a generic skill that virtually all faculty want students to develop, though the style of expression may vary from discipline to discipline. Consider, for example, the difference between writing about a scientific experiment and building an argument in political science. In

one of his *Assessment Measures* columns, Gary R. Pike describes some of the thinking that must form the basis for assessing writing. For instance, will students be given one opportunity to write an essay in a specified amount of time, or will they have several opportunities to draft and refine their product? Will the writing prompt ask students to develop a personal narrative or a persuasive essay? Will scoring be holistic or analytical?

Information literacy is a second broad set of skills that most faculty want students to strengthen while they are in college. My colleague, Howard R. Mzumara, and I describe a process for selecting instruments that may be used to assess information literacy and technological competence. Once again, a prudent first step is to define what campus faculty mean by the concept of information literacy. After that often-difficult discussion has taken place, colleagues can study the components of existing instruments to see if there is a good match with the campus definition or if a locally designed measure is needed.

A third skill that is a stated outcome of most approaches to general education is critical or analytical thinking. Reasoning and problem solving are other terms often linked with critical thinking. In her article, Diane Kelly-Riley recounts the story of a cross-disciplinary campus project that produced a seven-dimension rubric for teaching and assessing critical thinking in every course. Studies have demonstrated that integrating the *Critical Thinking Rubric* in classroom experiences improves critical thinking scores more than is the case when the same instructor teaches the same course without using the rubric.

Moral awareness is a concept not frequently identified as a specific goal of general education. But “a liberal education designed to develop versatile and critical thinkers who can adapt to the professional and ethical challenges they will confront” is important to the faculty who have designed the curriculum at the United States Military Academy, according to James JF Forest and Bruce Keith. In a developmental sequence that takes place throughout the student experience, cadets first learn to recognize moral issues and the ways in which ethics influence decision-making. Then they consider ethical options, counterarguments, and the implications of competing views. In later courses they must analyze cases, identify their moral considerations, and develop morally acceptable

responses. Data from multiple sources are used to assess cadets' moral awareness, including a survey of students' confidence in their achievement of the moral awareness goal, students' course products, and employer feedback drawn from focus group interviews of former battalion commanders.

The focus of the faculty-developed assessment approaches discussed to this point has been direct measurement of student performance. Forrest and Keith introduce an indirect measure—a student survey in which cadets reflect on their experiences and report their perceptions of the quality of their own learning. Anne Hummer also describes an indirect measure. Dissatisfied with the usual methods—portfolios and performance assessment via video-taping—of measuring multifaceted communication skills, Hummer's colleagues developed the *Student Perception of Communication Skills*. This 50-item questionnaire asks students about self-development of oral and written communication as well as nonverbal and listening skills. Questions also cover peer, faculty, and course contributions to the development of these skills. Hummer's essay addresses the importance of investigating the technical qualities of assessment instruments. She and her colleagues studied the internal consistency and the factor structure of their instrument. Hummer concludes with a detailed summary of the changes faculty made in their teaching based on areas of weakness revealed in students' questionnaire responses.

Valuing diversity is the component of the community college curriculum on which Scott Hunt focuses his selection in this issue. Freshman Experience and Capstone courses provide opportunities for administration of pretest and posttest versions, respectively, of a scenario that gauges students' cultural awareness and appreciation of diversity.

Assessment Methods Applicable Across Knowledge and Skills Areas

We shift our focus now from locally developed assessment approaches that have been applied in specific knowledge and skills areas to generic methods that may be utilized in a variety of these areas. Capstone courses,

mentioned here first by Hunt in the previous selection, constitute one such generic approach that is applicable in virtually every discipline. As Terrel L. Rhodes and Susan Agre-Kippenhan note in their article, capstones provide a rich context indeed for administering multiple assessment techniques. In the capstone courses these authors describe, students apply what they have learned in their majors to address community challenges. Student focus groups, reflective essays, end-of-course evaluations, and a post-capstone student survey are some of the ways in which student learning is assessed in these capstones.

Classroom assessment is another effective methodology that can be applied across disciplines. Barbara E. Walvoord, Barbara Bardes, and Janice Denton recount some of the faculty objections to the use of standardized tests that were outlined in previous selections, then tell us how faculty can apply locally developed rubrics to student work in their own classrooms. As these authors note, faculty are much more likely to “close the loop” by using assessment data to improve student learning if they control the process of assessment in their individual classrooms. As faculty uncover student weaknesses and modify their teaching to address these, departmental colleagues can benefit from this information and share some of their own wisdom derived from classroom assessment. Department faculty can share their findings at college and institutional levels. Thus classroom assessment can form the basis for assessing unit and institutional effectiveness as well as the generic skills of individuals.

Phil Speary tells us about a similar classroom-based assessment process adapted for use at a community college. Speary’s colleagues assess such skills as critical thinking, speaking, listening, and teamwork in their own courses, using their own assignments. Next they apply a standardized rubric for each skill to their students’ classroom products. The new wrinkle here is that Speary’s Office of Assessment aggregates data from the rubrics centrally and reports scores to a central committee that can suggest warranted changes in curriculum, instruction, and student services across the institution.

The late Donald W. Farmer was instrumental in establishing outcomes assessment as a distinguishing characteristic of education at King’s

College, where “students engage in multiple performance-based assessment experiences in the classroom from the point of entry to the point of graduation.” Assessment strategies are embedded in each course and each student has a Competence Growth Plan for eight “transferable skills of liberal learning.” Pre- and post-assessments occur in each course, and assessment of the eight transferable skills is the responsibility of each discipline. A unique strategy is the Sophomore-Junior Diagnostic Project that students begin in a sophomore class, carry out over the summer, and present in a junior-level course. Collective consideration of evidence derived from assessment educates faculty about the effectiveness of their work with students. King’s faculty also use assessment as a basis for their scholarship. Conference programs and scholarly journals contain a noteworthy number of presentations by King’s faculty.

Our final selection, by Philip I. Kramer, reports on an effort in Utah to extend a common approach to assessment across the public institutions in an entire state. In one sense we end this collection where we began—discussing another failed attempt to use a standardized test of generic skills—in this case the Collegiate Assessment of Academic Proficiency (CAAP)—to compare institutions with diverse missions. But in Utah, as in a number of other states, policymakers agreed to authorize a representative group of faculty from various colleges and universities to design their own “content-embedded assessment instruments.” The faculty group completed the difficult work of agreeing on nine components of general education that would apply across the system. Then faculty in four disciplines—economics, history, mathematics, and political sciences—developed banks of multiple choice, true-false, and fill-in-the-blank questions for use in a pretest-posttest design intended to measure value added. Faculty were permitted to choose their own items from the banks of test questions. Where the pre- and posttests were applied, students made dramatic gains. But no one was satisfied with the outcomes of the assessment process. Since there was no common test, institutions could not be compared, so policymakers were disappointed. A host of unaddressed methodological issues left faculty wondering how the test results could be used.

The Utah case study demonstrates yet again the enormous set of difficulties we face in attempting to design measures of the learning outcomes associated with general education that will satisfy policy-makers and other stakeholders in higher education. The preponderance of articles selected for this issue demonstrate that at many colleges and universities across the country, faculty are developing course- and curriculum-based measures that enable them to detect strengths and weaknesses in the learning of groups of students. This evidence is being used to improve courses, curricula, and student services. But in developing the measures that our stakeholders envision for the purpose of comparing the quality and accountability of various institutions, we have made very little progress in the two decades since Robert Thorndike conducted his studies in the mid-1980s.