A Brief History of Student Learning Assessment

How We Got Where We Are and a Proposal for Where to Go Next

Richard J. Shavelson

With a Foreword by
Carol Geary Schneider
and Lee S. Shulman

the
Academy
in
Transition

A Brief History of
Student Learning
Assessment

How We Got Where We Are and
a Proposal for Where to Go Next

Richard J. Shavelson

With a Foreword by
Carol Geary Schneider
and Lee S. Shulman

Contents

the Academy *in* Transition

Other titles in the series:

For information about these and other AAC&U publications, or to place an order, visit www.aacu.org, e-mail pub_desk@aacu.org, or call 202-381-3760.

About This Series

The Association of American Colleges and Universities (AAC&U) has a long history of working with college leaders across the country to articulate the aims of a liberal education in our time. AAC&U is distinctive as a higher education association. Its mission focuses centrally on the quality of student learning and the changing purpose and nature of undergraduate curricula.

AAC&U has taken the lead in encouraging and facilitating dialogue on issues of importance to the higher education community for many years. Through a series of publications called the Academy in Transition—launched in 1998 with the much-acclaimed *Contemporary Understandings of Liberal Education*—AAC&U has helped fuel dialogue on such issues as the globalization of undergraduate curricula, the growth of interdisciplinary studies, how liberal education has changed historically, and the increase of college-level learning in high school. The purpose of the series, which now includes ten titles, is to analyze changes taking place in key areas of undergraduate education and to provide "road maps" illustrating the directions and destinations of the changing academy.

During transitions, it is important to understand context and history and to retain central values, even as the forms and structures that have supported those values may have to be adapted to new circumstances. For instance, AAC&U is convinced that a practical and engaged liberal education is a sound vision for the new academy, even if the meanings and practices of liberal education are in the process of being altered by changing conditions. As the titles in this series suggest, AAC&U's vision encompasses a high-quality liberal education for all students that emphasizes connections between academic disciplines and practical and theoretical knowledge, prizes general education as central to an educated person, and includes global and cross-cultural knowledge and perspectives. Collectively, the papers published in the Academy in Transition series point to a more purposeful, robust, and efficient academy that is now in the process of being created. They also encourage thoughtful, historically informed dialogue about the future of the academy.

AAC&U encourages faculty members, academic leaders, and all those who care about the future of our colleges and universities to use these papers as points of departure for their own analyses of the directions of educational change. We hope this series will encourage academics to think broadly and creatively about the educational communities we inherit, and, by our contributions, the educational communities we want to create.

Foreword

Many of us in higher education today are thinking hard about assessment. But often our cogitations tend toward what can only be called wishful thinking.

A wish that has no doubt washed over all of us at one time or another is that if we simply ignore assessment, or hold it off long enough, the issues (like the misguided new administrator) will finally give up and go away. But in our wiser moments, we know that this is not an answer. Indeed, as is clear from Richard Shavelson's lively tour of the twists and turns of assessment over the past century, the names may change, and the technology has evolved, but assessment has stayed with us with great persistence. "Today's demand for a culture of evidence of student learning appears to be new" Shavelson tells us, but it turns out "to be very old," and there's no wishing it away. Moreover, we should not be eager to wish it away.

Nor is there a magic bullet. 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—the test with psychometric properties so outstanding that we can base high-stakes decisions on the results of performance on that measure alone.

This wish is not only self-indulgent but self-defeating. Ironically, no test can possess such properties because, to achieve validity, test designers have to narrow the focus on any particular instrument to a sobering degree. Thus, the better the arguments we can make regarding the validity of any given measure—whether of knowledge, skills, or some other virtue—the less appropriate that measure is as the basis for consequential decisions about a student's overall learning gains, much less as the sole determinant of an institution's educational quality.

Thinking of assessment as primarily a technical challenge—though certainly it *is* that, as Shavelson's analysis also makes clear—is another form of wishful thinking. The far-reaching questions raised through assessment cannot be solved through technical ingenuity alone.

What's needed, of course, is *educational* thinking, and happily there has been a good deal of that in the past two decades of assessment activity. With the wave of state mandates for assessment in the mid and later 1980s, and new accreditation requirements in the 1990s, campuses began to organize themselves to respond. Many did so grudgingly, and there were plenty of missteps, misunderstandings, and dead ends. But there were also wonderful examples of what can happen when educators take up the challenge to figure out and clearly articulate what they want their students to know and be able to do: the core task of assessment. Many of these efforts were

funded by the U.S. Department of Education's Fund for the Improvement of Postsecondary Education, and the results, in turn, provided models and momentum for additional campuses that came together—hundreds of them—at the assessment forum led by the American Association for Higher Education until 2005, when that organization closed its doors.

Toward the end of his essay, Shavelson makes a crucial point that campuses committed to assessment know well: *that assessment all by itself is an insufficient condition for powerful learning and improvement.* Of course, more and better evidence of student learning is important, but knowing what to *make* of that evidence, and how to act on it, means getting down to core questions about the character of the educational experience and the goals of liberal learning. These are not questions that higher education can dare leave to the testing companies or to external agencies, no matter how well intentioned and enlightened.

This view of assessment has become central to the work of both the Carnegie Foundation for the Advancement of Teaching and the Association of American Colleges and Universities (AAC&U). Shavelson traces Carnegie's work over the first part of the twentieth century as a story about standards and standardization. But the needs are different today, and Carnegie's more recent work places much greater emphasis on the role of faculty in exploring what our students do—and don't—learn. The foundation's extensive work on the scholarship of teaching and learning, for instance, has helped fuel a movement in which "regular" faculty, across the full spectrum of disciplines and institutional types, are treating their classrooms and programs as laboratories for studying student learning in order to improve it. Seen through the lens of classroom inquiry, assessment is a feature of *the pedagogical imperative* in which faculty see themselves as responsible for the learning of their students and for deepening our collective sense of the conditions in which important forms of learning can occur.

Through its Liberal Education and America's Promise (LEAP) initiative, AAC&U is working with its member campuses to develop assessments that strengthen students' learning and assess their best work rather than just the attainment of a few narrowly defined foundational skills and/or basic knowledge. As AAC&U's board of directors put it in their official statement on assessment (2005, 3), colleges and universities should hold themselves "accountable for assessing [their] students' best work, not generic skills and not introductory levels of learning."

In its recently released LEAP report, *College Learning for the New Global Century*, AAC&U recommends tying assessment efforts much more closely to the curriculum and to faculty priorities for student learning across the curriculum. The report affirms, as well, that any national assessment measure, however well developed, is only part of the solution to the

problem of underachievement. As the report notes, "standardized tests that stand outside the regular curriculum are, at best, a weak prompt to needed improvement in teaching, learning, and curriculum. Tests can, perhaps, signal a problem, but the test scores themselves do not necessarily point to where or why the problem exists or offer particulars as to solutions" (2007, 40).

A fuller strategy, the LEAP report proposes, would prepare students to produce a substantial body of work—capstone projects and/or portfolios—that require their best efforts. The resulting accomplishments should be assessed for evidence of students' competence on liberal education outcomes such as analytical reasoning and integrative learning, as well as their achievement in their chosen fields. Standardized assessments can then fill out the emerging picture, providing the ability to benchmark accomplishment against peer institutions, at least on some aspects of student learning.

As the LEAP report notes, "however the assessments are constructed . . . the framework for accountability should be students' ability to apply their learning to complex problems. Standards for students' expected level of achievement also will vary by field, but they should all include specific attention to the quality of the students' knowledge, their mastery of key skills, their attentiveness to issues of ethical and social responsibility, and their facility in integrating different parts of their learning" (2007, 41–2).

Richard Shavelson offers an important historical context to consider as institutions across the country continue to develop new methods of assessment in response to renewed calls for greater accountability and, more importantly, the urgent need to raise levels of student achievement. He helps us better understand the "state-of-the-art" in standardized testing today, and what we should ask from testing agencies in the future. Above all, he helps us understand why psychometricians themselves are so opposed to any efforts at institutional ranking or comparisons based on standardized tests.

We are grateful to Richard Shavelson for taking the time to put current debates in a larger historical and educational forum. Everyone who is thinking today about assessment and public accountability will benefit greatly from the insights this study provides.

Carol Geary Schneider
President, Association of American Colleges and Universities
Lee S. Shulman
President, the Carnegie Foundation for the Advancement of Teaching

I. Introduction

Over the past thirty-five years, state and federal policy makers, as well as the general public, have increasingly been pressuring higher education to account for student learning and to create *a culture of evidence*. While virtually all states already use proxies (e.g., graduation rates) to report on student performance, states are now being pressured to measure learning directly. U.S. Secretary of Education Margaret Spellings' Commission on the Future of Higher Education, for example, has called for standardized tests of students' critical thinking, problem solving, and communication skills.

While the current demand to establish a culture of evidence appears to be new, it has a long lineage. The future development of this culture may very well depend on how well we appreciate the past. *Cultures of evidence will not automatically lead to educational improvement, if what counts as evidence does not count as education.* Narrow definitions and narrow tests of what count as learning outcomes in college may very well distort the culture of evidence we seek to establish. As we shall see from the past, and as we know from current studies (Immerwahr 2000; AAC&U 2007), there is more to be learned and assessed in higher education than the broad abilities singled out by the Spellings Commission for measurement by standardized tests. These additional outcomes include learning to know, understand, and reason in an academic discipline. They also include personal, civic, moral, social, and intercultural knowledge and actions—outcomes the Educational Testing Service has described as "soft." Some experts say that such "soft" outcomes cannot be measured adequately because "the present state of the art in assessing these skills is not adequate for supporting the institution of a nationwide set of standardized measures" (Dwyer, Millett, and Payne 2006, 20). But this position is unsatisfactory. This set of outcomes—which, following the lead of the Association of American Colleges and Universities, I will call personal and social responsibility (PSR) skills—are every bit as demanding as the academic skills that often get labeled exclusively as *the cognitive skills* and are too important not to be measured. *If we do not measure PSR skills, they will drop from sight as accountability pressures force campuses to focus on a more restricted subset of learning outputs that can be more easily and less expensively measured.*

The outcomes framework depicted in figure 1 demonstrates the importance of extending the range of outcomes we assess beyond broad abilities. Such outcomes could range from the development of specific factual, procedural, and conceptual knowledge and reasoning in a discipline

(such as history) to the development of the skills on which the Spellings Commission focused (critical thinking, problem solving, and communication), to the development of reasoning applicable to a very wide variety of situations, or to the development of intelligence. Moreover, "cognitive" outcomes include PSR skills insofar as reasoning and thinking are involved in personal relations, moral challenges, and civic engagement. The PSR skills are not so soft; they involve cognition and more, *as do academic skills*. Finally, the arrows in figure 1 remind us that general abilities influence the acquisition of knowledge in concrete learning environments, that direct experiences are the stuff on which reasoning and abstract abilities are developed, and that cognitive performance on academic and PSR skills is influenced by the interaction of individuals' accumulated experience in multiple environments with their inheritance.

Furthermore, the standardized tests that the Spellings Commission and others have in mind for outcomes assessment are not interchangeable. There are different ways to measure student learning; some standardized tests focus only on a narrow slice of achievement, while others focus on broader abilities developed over an extended course of study. Especially for

Abstract, Process Oriented

Intelligence

General   Fluid   Crystallized

Inheritance x Accumulated Experience

General Reasoning

Verbal   Quantitative   Spatial

Example: Graduate Record Examination

Broad Abilities

Reasoning   Critical Thinking   Problem Solving   Decision Making

Communicating
In Broad domains

(Disciplinary—Humanities, Social Sciences, Sciences—
And Responsibility—Personal, Social, Moral, and Civic)
Example: Collegiate Learning Assessment

Concrete Context-Oriented

Knowledge, Understanding and Reasoning

In Major Fields and Professions (American Literature, Business)

Example: ETS' Major Field Tests
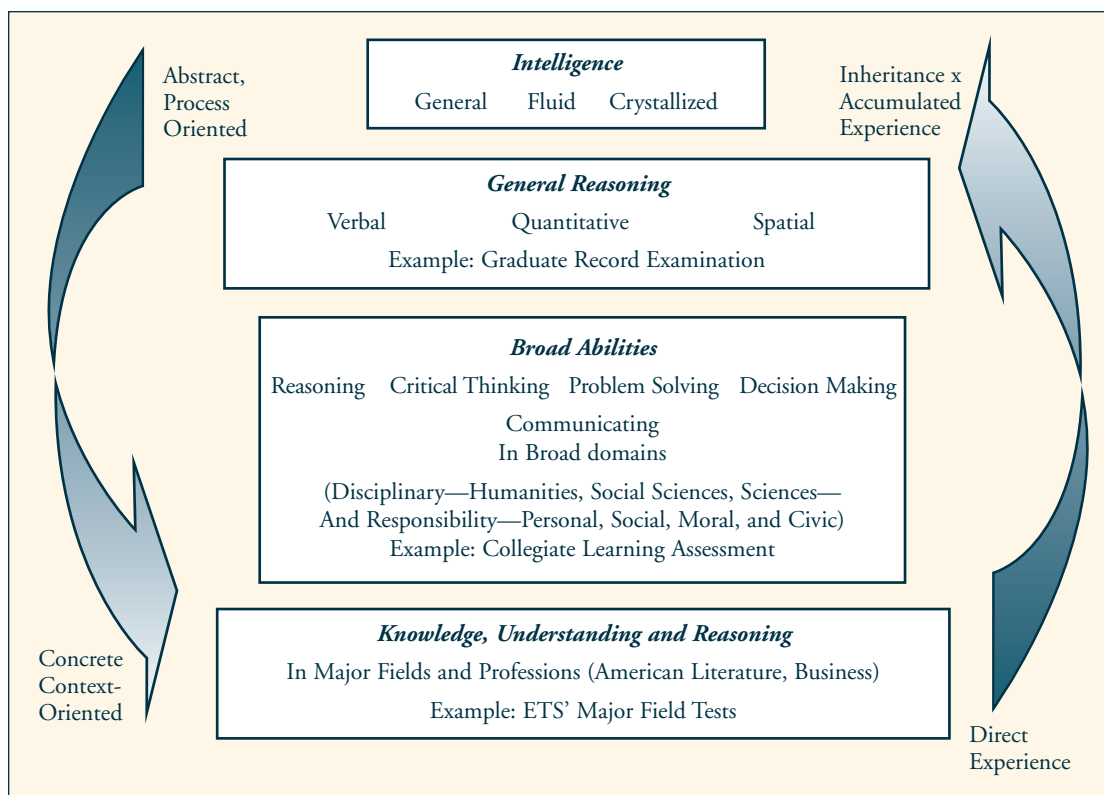
Direct Experience

Figure 1. Framework for student learning outcomes. (Adapted from Shavelson and Huang 2003, 14.)

higher education, the different assumptions about what ought to be measured that are embedded in every assessment instrument need to be clarified and carefully considered before specific tests are chosen to assess students' cumulative gains from college study.

The multiple-choice technology developed almost a hundred years ago, for example, is inherently limited when it comes to measuring the full array of student learning outcomes depicted in figure l. Multiple-choice measures have a long history, as we shall see. They are the basis of the standardized tests that are often used today, including the Collegiate Assessment of Academic Proficiency (CAAP), the Measure of Academic Proficiency and Progress (MAPP), and the College Basic Academic Subjects Examination (CBASE). The MAPP was recommended by the Spellings Commission as ways of assessing student learning in college. But these measures are limited in their ability to get at some of the more complex forms of reasoning and problem solving that are commonly viewed as distinctive strengths of American higher education.

If the learning outcomes of higher education are narrowly measured because cost, capacity, and convenience dictate reductive choices, then we stand the risk of narrowing the mission and diversity of the American system of higher education, as well as the subject matter taught. What we need to do instead is to learn from the rich history of student learning assessment and take responsible steps to develop and measure the learning outcomes our nation values so highly.

II. A Brief History of Learning Assessment

Surprisingly, our journey begins with the Carnegie Foundation for the Advancement of Teaching, which is so well known for the "Carnegie unit" and TIAA. The foundation was arguably the ringleader of student learning assessment. Howard Savage (1953), a staff member and historian of the foundation in its early days, attributes Carnegie's leadership in college learning assessment to its first president, Henry Pritchett, who was motivated by his concern for the quality of higher education and his recognition of the potential impact that the emergence of "objective testing" might have on monitoring that quality. Walter A. Jessup, the foundation's third president, later put what had become the foundation's vision this way:

> The central problems [in improving higher education] are three in number: first, the setting up of generally accepted standards of achievement; secondly, the devising of methods of measuring this achievement and holding pupils to performance; and thirdly, the introduction of such flexibility in educational offerings that each individual may receive the education from which he is able to derive the greatest benefit. (Kandell 1936, vii)

Pritchett's passion was shared by his chief staff member, William S. Learned, "a man who had clear and certain opinions about what education ought to be . . . [with] transmission of knowledge as the *sine qua non*" (Lagemann 1983, 101). Learned became the instrument through which the foundation transformed higher education learning assessment.[1] Together with Columbia College's Ben D. Wood, who held the view "that thinking was dependent upon knowledge and knowledge dependent upon facts" (Lagemann 1983, 104), Learned led a large-scale assessment of college learning in the state of Pennsylvania. Learned parlayed this experience into the development of the Graduate Record Examination and germinated the idea of a "National Examination Board," a national testing agency that, twenty years later, became the Educational Testing Service.

The assessment of college learning[2] evolved through four eras: (1) the origin of standardized tests of learning: 1900–1933; (2) the assessment of learning for general and graduate education: 1933–47; (3) the rise of test providers: 1948–78; and (4) the era of external accountability: 1979–present. For ease of reference, the tests and testing programs discussed below are summarized in table form in the appendix.

The Origin of Standardized Tests of Learning: 1900–1933

The first third of the twentieth century marked the beginning of the use of standardized, objective testing to measure learning in higher education. The Carnegie Foundation led the movement; in 1916, William Learned tested students "in the experimental school at the University of Missouri in arithmetic, spelling, penmanship, reading, and English composition using recognized tests, procedures, and scales, and a statistical treatment that though comparatively crude was indicative" (Savage 1953, 284). E. L. Thorndike's study of engineering students followed. Thorndike tested students at the Massachusetts Institute of Technology, the University of Cincinnati, and Columbia University on "all or parts of several objective tests in mathematics, English and physics" (Savage 1953, 285). These tests focused on content knowledge, largely tapping facts and concepts (declarative knowledge) and mathematical routines (procedural knowledge). The early tests were "objective"; students responded by selecting an answer where one answer was correct. Compared to the widely used essay examination, these tests gained reliability in scoring and content coverage per unit of time.

The monumental Pennsylvania Study, conducted between 1928 and 1932, emerged from this start. It tested thousands of high school seniors, college students, and even some college faculty members using extensive objective tests of largely declarative and procedural content knowledge. In many ways, the Pennsylvania Study was exemplary; it proceeded with a clear conception of what students should achieve and how learning should be measured. In other ways, however, it reflected its time; the study focused on knowledge and required compliant students to sit for hours of testing.[3]

In the 1928 pilot study, no less than 70 percent of all Pennsylvania college seniors, or 4,580 students, took the assessment as did about 75 percent of high school seniors, or 26,500 students. Of the high school seniors, 3,859 entered a cooperating Pennsylvania college; 2,355 remained through their sophomore year in college, and 1,187 remained through their senior year (Learned and Wood 1938, 211).

The assessment itself was a whopping twelve hours and 3,200 items long. (The examiners expressed regret at not being more comprehensive in scope!) Comprised of selected-response questions—for example, multiple-choice, matching, and true-false—the assessment covered nearly all areas of the college curriculum. The main study focused on student *learning*, not simply on achievement in the senior year, by testing students during their senior year of high school and then testing them again during their sophomore and senior years in college.

The Pennsylvania Study is noteworthy because it laid out a conception of what was meant by undergraduate achievement *and* learning, assuming that achievement was the result of college learning defined as the accumulation of breadth and depth of content knowledge. It also focused heavily and *comprehensively* at the knowledge level, especially on declarative and procedural knowledge. Nevertheless, because it included an intelligence test, the assessment program tapped the extremes of the outcomes framework: content knowledge and general intelligence. Moreover, the Pennsylvania Study employed technology for assessing learning and achievement—objective testing—that followed directly from the study's conception of learning. If knowledge were understood as the accumulation of learning content, then objective testing could efficiently verify—literally index—the accumulation of that knowledge (Learned and Wood 1938, 372). Finally, the Pennsylvania Study is also noteworthy because, unlike assessments done today, it collected data in designs that provided evidence of both achievement and learning. In some cases, the comparison was across student cohorts, or "cross-sectional," including high school seniors, college sophomores, and college seniors. In other cases, it was longitudinal; the same high school seniors tested in 1928 were tested again as college sophomores in 1930 and then as seniors in 1932.

The Assessment of Learning in General and Graduate Education: 1933–47

This era saw the development of both general education and general colleges in universities across the country, as well as the evolution of the Graduate Record Examination (GRE). The Pennsylvania Study had provided an existence proof; comprehensive assessment of student learning was feasible. Individual institutions, as well as consortia, put together test batteries designed primarily to assess cognitive achievement. Perhaps most noteworthy in this progressive-education period, with its focus on the whole student, was the attempt to measure not only cognitive outcomes across the spectrum but also the personal, social, and moral outcomes of general education.

Here, I briefly treat learning assessment in general education because, as it emerged in some important cases, it diverged from rather than adopted the Carnegie Foundation's view of education and learning assessment. The University of Chicago's approach presages contemporary developments; the Cooperative Study presages the call for "soft-skills." I then focus attention on the evolution of the GRE.

General Education and General Colleges. The most notable examples of general education learning assessment in this era are the University of Chicago College program and the Cooperative Study of General Education (for additional programs, see Shavelson and Huang 2003). The former reflected thinking in the progressive era, while the latter had its roots in the Carnegie Foundation's conception of learning but also embraced progressive notions of human development as well.

In the Chicago program, a central university examiner's office, rather than individual faculty in their courses, was responsible for developing, administering, and scoring tests of student achievement in the university's general education program (Frodin 1950). Whereas the Pennsylvania Study assessed declarative and procedural knowledge, the Chicago examinations tested a much broader range of knowledge and abilities: the use of knowledge in a variety of unfamiliar situations; the ability to apply principles to explain phenomenon; and the ability to predict outcomes, determine courses of action, and interpret works of art. The Chicago comprehensive exams were characterized by open-ended essays and multiple-choice questions demanding interpretation, synthesis, and application of new texts (primary sources).[4]

The Cooperative Study of General Education, conducted by a consortium of higher education institutions, stands out from assessment initiatives at individual campuses. The participating institutions believed they would benefit from a cooperative approach to the improvement of general education (Executive Committee of the Cooperative Study in General Education 1947; Dunkel 1947; Levi 1948). To that end, and in order to assess students' achievement *and well-being*, the consortium developed the Inventory of General Goals in Life, the Inventory of Satisfactions Found in Reading Fiction, the Inventory of Social Understanding, and the Health Inventories.

The Evolution of the Graduate Record Examination: From Content to General Reasoning. While learning assessment was in full swing, Learned and Wood parlayed their experience with the Pennsylvania Study into an assessment for graduate education. In proposing the "Co-operative Graduate Testing Program," Learned noted that, with increased demand for graduate education following the Depression, the A.B. degree had "ceased to draw the line between the fit and the unfit" (Savage 1953, 288). Graduate admissions and quality decisions needed to be based on something more than the number of college credits.

In October 1937, Learned's team worked with the graduate schools at Columbia, Harvard, Princeton, and Yale to administer seven tests designed to index the quality of students in graduate education. This was the first administration of what was to be the Graduate

Record Examination (GRE). The program was a success and grew by leaps and bounds (see fig. 2). And at a time when the Carnegie Foundation was struggling to keep its faculty retirement system (TIAA) afloat, it was also a growing financial and logistic burden. Ultimately, the foundation was motivated by these stresses to pursue the establishment of an independent, national testing service.

Like the examinations used in the Pennsylvania Study, the original GRE was a comprehensive and objective test focused largely on students' content knowledge, but it also tapped verbal reasoning and was used to infer students' fitness for graduate study (Savage 1953). In 1936, a set of "profile" tests was developed to cover the *content* areas of a typical undergraduate general education program. To be completed in two half-day sessions totaling six hours, the tests measured knowledge in mathematics, the physical sciences, social studies, literature and fine arts, and one foreign language. The verbal factor was "developed primarily as a measure of ability to discriminate word meanings" (Lannholm and Schrader 1951, 7). In 1939, sixteen Advanced Tests in subject major fields were added to the GRE, and in 1949, a general education section was added to the Profile Tests in order to tap "effectiveness of expression" and to provide a "general education index" (see ETS 1953).

The fall of 1949 saw a landmark in student learning assessment: in a shift from testing content to testing general reasoning, ETS introduced a GRE Aptitude Test with the kind of verbal and quantitative sections we see today. Then, in 1952, it introduced the now standard
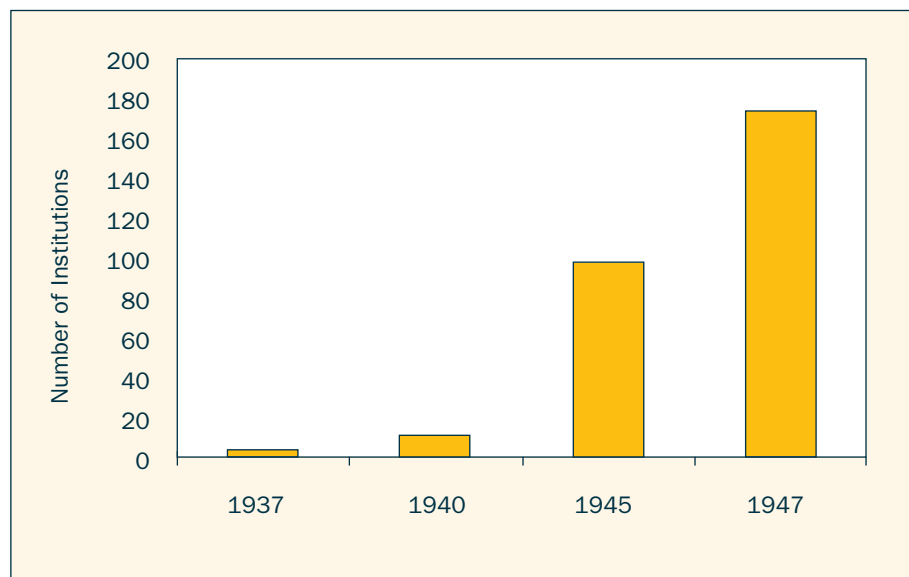

Figure 2. GRE growth over its first ten years.

scale for reporting scores (the normal distribution with mean 500 and standard deviation 100). In 1954, ETS continued the shift away from content and toward general reasoning by replacing both the Profile Tests and the Tests of General Education with the "Area Tests," which served as a means of assessing broad outcomes of the liberal arts. The Area Tests focused on academic majors in the social and natural sciences and the humanities. They emphasized reading comprehension, understanding, and interpretation, often providing requisite content knowledge "because of the differences among institutions with regard to curriculum and the differences among students with regard to specific course selection" (ETS 1966, 3).

The Rise of the Test Providers: 1948–78

During the period following World War II, with funding from the G.I. Bill of Rights, postsecondary education enrollments mushroomed, as did the number of colleges to accommodate the veterans *and* the number of testing companies to assist colleges in screening them—most notably ETS, created in 1948, and the American College Testing (ACT) program, created in 1959.

Tests Provided by Testing Organizations to Assess Student Learning. By the time the Carnegie Foundation had transferred the GRE to ETS, completing its move out of the testing business, it had left an extraordinarily strong legacy of objective, group-administered, cost-efficient testing using selected response questions—now solely multiple-choice. That legacy has endured into the twenty-first century. The precursors of today's major learning assessment programs were developed by testing organizations in this era (Shavelson and Huang 2003, 2006). These 1960s and 1970s testing programs included ETS's Undergraduate Assessment Program, which incorporated the GRE, and ACT's College Outcomes Measures Project (COMP). The former evolved via the Academic Profile into today's Measure of Academic Proficiency and Progress (MAPP), and the latter evolved into today's College Assessment of Academic Proficiency (CAAP).

However, several developments in the late 1970s, reminiscent of the progressive era, augured for a change in the course set by Learned and Wood. Faculty members were not entirely happy with multiple-choice tests. They wanted to get at broader abilities—such as the ability to communicate, think analytically, and solve problems—in a holistic manner. This led to several new developments. ETS studied constructed-response tests that tapped communication skills, analytic thinking, synthesizing ability, and social/cultural awareness (Warren 1978).

ACT experimented with open-ended performance-based assessments that sought to measure skills for effective functioning in adult life in social institutions, in using science and technology, and in using the arts. And the state of New Jersey developed Tasks in Critical Thinking, which sampled real-world tasks in a "performance-based assessment … [that measured] the ability to use the skills of inquiry, analysis, and communication" with prompts that "do not assess content or recall knowledge" (ETS 1994, 2). These assessment programs were designed to embrace what college faculty considered to be important learning outcomes.

For a short period, these learning assessments set the mold. But due to time and cost limitations, as well as difficulties in securing and training people to score responses and in achieving adequate reliability, they either faded into distant memory or morphed back into multiple-choice tests. For example, the COMP began as a pathbreaking performance-based assessment. Its content was sampled from materials culled from everyday experience including film excerpts, taped discussions, advertisements, music recordings, stories, and newspaper articles. The test sought to measure three process skills—communicating, solving problems, and clarifying values—in a variety of item formats, including multiple-choice, short answer, essay, and oral response (an atypical format). COMP, then, bucked the trend toward multiple-choice tests of general abilities by directly observing performance sampled from real-world situations.

The test, however, was costly in terms of time and scoring. Students were given six hours to complete it in the 1977 field trials; the testing time was reduced to four and a half hours in the 1989 version. Raters were required to score much of the examination. As both a consequence and a characteristic of trends, a simplified "Overall COMP" was developed as a multiple-choice-only test. In little more than a decade, however, this highly innovative assessment was discontinued altogether due to the costliness of administration and scoring. Roughly the same story describes the fate of Tasks in Critical Thinking (see Erwin and Sebrell 2003).[5]

The influence of the Carnegie Foundation, then, waned in the mid-1970s. However, as we shall see, the foundation's vision of objective, selected-response testing continued to influence the standardized learning assessment programs of ETS, ACT, and others.

The Era of External Accountability: 1979–Present

By the end of the 1970s, political pressures to assess student learning and hold campuses accountable had coalesced. While in the 1980s only a handful of states had some form of mandatory standardized testing (e.g., Florida, Tennessee), public and political demand for

such testing increased into the new millennium (Ewell 2001). To meet this demand, some states (e.g., Missouri) created incentives for campuses to assess learning, and campuses responded by creating learning assessment programs.

Tests of College Learning. ETS, ACT, and others were there to provide tests. Indeed, a wide array of college learning assessments following in the tradition of the Carnegie Foundation was available. Currently, ETS provides the MAPP, ACT provides CAAP, and the College Resource Center at the University of Missouri, Columbia, offers the College Basic Academic Subjects Examination (CBASE). All are multiple choice test batteries. MAPP measures college-level reading, mathematics, writing, and critical thinking in the context of the humanities, social sciences, and natural sciences to enable colleges and universities to improve their general education outcomes. CAAP measures reading, writing, mathematics, science, and critical thinking to enable postsecondary institutions to evaluate and enhance general education programs. CBASE is a criterion-referenced achievement examination of English, mathematics, science, and social studies that serves both to qualify individuals for entry into teacher education programs and to test general academic knowledge and skills.

Vision for Assessing Student Learning. As we saw, at the end of the 1970s, objective testing was incompatible with the way faculty members either assessed student learning or wanted student learning to be assessed. For them, *life is not a multiple-choice test*. Rather, faculty members like the open-ended, holistic, problem-based assessments exemplified by, for example, Tasks in Critical Thinking. Intuitively, faculty members suspected that the kind of thinking stimulated and performance assessed by multiple-choice and other highly structured tests is different from that stimulated and assessed by more open-ended tasks. And empirical evidence supports their intuition.

While a multiple-choice test and a "constructed-response" test may produce scores that are correlated with each other, this correlation does not mean that the same kind of thinking and reasoning is involved (Martinez 1999; National Research Council 2001). Student performance varies considerably depending upon whether a task is presented as a multiple-choice question, an open-ended question, or a concrete performance task (Baxter and Shavelson 1994). For example, Lythcott (1990, 248) found that "it is possible, though not our intention, for [high school and college] students to produce right answers to chemistry problems without really understanding much of the chemistry involved." Moreover, Baxter and Shavelson (1994) found that middle school students who solved electric circuit problems hands-on could not solve the same problems represented abstractly in a multiple-choice test; these students did not

make the same assumptions that the test developers made. Finally, using a "think aloud" method to tap into students' cognitive processing, Ruiz-Primo and colleagues (2001) found that students reasoned differently on highly structured assessments than on loosely structured assessments. In the former case students "strategized" as to what alternative fit best, while in the latter they reasoned through the problem.

To illustrate the difference between multiple-choice and open-ended assessments, consider the following concrete example from the Collegiate Learning Assessment (CLA). College students are asked to assume that they work for "DynaTech"—a company that produces industrial instruments—and that their boss has asked them to evaluate the pros and cons of purchasing a "SwiftAir 235" for the company. Concern about such a purchase has risen with the report of a recent SwiftAir 235 accident. When provided with an "in-basket" of information, some students, quite perceptively, recognize that there might be undesirable fallout if DynaTech's own airplane crashed while flying with DynaTech instruments. Students are not prompted to discuss such implications; they have to recognize these consequences on their own. There is no way such insights could be picked up by a multiple-choice question.

Finally, consistent with the views of faculty, both the American Association of State Colleges and Universities (AASCU) and members of the Spellings Commission have in mind a particular standardized learning assessment: the CLA.

The best example of direct value-added assessment is the Collegiate Learning Assessment (CLA), an outgrowth of RAND's Value Added Assessment Initiative (VAAI) that has been available to colleges and universities since spring 2004. The test goes beyond a multiple-choice format and poses real-world performance tasks that require students to analyze complex material and provide written responses (such as preparing a memo or policy recommendation). (AASCU 2006, 4)

This brief history has now arrived at the present. In contrast to the evolution in multiple-choice testing technology, the Council for Aid to Education has taken Tasks in Critical Thinking and COMP to what might be considered the next level by marrying the open-ended assessment of real-world, holistic tasks and the use of computer technology to assess ability and learning.[6] A closer look at the CLA may provide insight into a next generation of learning assessments.

III. The Collegiate Learning Assessment

Just as new technology, objective testing, and computer scoring of response sheets (developed for Learned by IBM) revolutionized learning assessment at the turn of the twentieth century, so too have new information technology and statistical sampling ushered in a significant change in college learning assessment at the turn of the twenty-first century. And yet, in some ways, the use of "new" assessment technology actually marks a return to the past. It represents a move away from selected-response, multiple-choice tests in favor of the concrete, complex, open-ended tasks embodied in assessments like the Collegiate Learning Assessment (CLA).

The roots of the CLA can be traced to progressive notions of learning that focus on critical thinking, analytic reasoning, problem solving, and written communication (see table 1). These capabilities are tapped in realistic "work-sample" tasks drawn from education, work,

Table 1. Characteristics of the Collegiate Learning Assessment

Characteristic	Attributes
Open-ended Tasks	• Tap critical thinking, analytic reasoning, problem solving, and written communication • Realistic work samples • Engaging tasks (as suggested by alluring titles such as "brain boost," "catfish," "lakes to rivers") • Applicable to different academic majors
Computer Technology	• Interactive Internet platform • Paperless administration • Natural language processing software for scoring students' written communication • Online rater scoring and calibration of performance tasks • Report institution's (and subdivision's) performance, as well as individual student performance (confidentially to the student)
Focus	• Institution or school/department/program within institutions • Not on individual student performance (although their performance is reported to them confidentially)
Sampling	• Samples students so that not all students perform all tasks • Samples tasks for random subsets of students • Creates scores at institution or subdivision/program level as desired (depending on sample sizes)
Reporting	• Controls for students' ability so that "similarly situated" benchmark campuses can be compared • Provides value added estimates—from freshman to senior year or with measures on a sample of freshmen and seniors • Provides percentiles • Provides benchmark institutions

and everyday issues. Such tasks are accessible to students from the wide diversity of majors and general education programs. Recent developments in information technology have enabled these rich tasks to be provided without overburdening students. The assessment is delivered on an interactive Internet platform that produces a paperless, electronic administration and online reports of results. Written communication tasks are scored using natural language processing software, and performance tasks are currently scored online by human raters whose scoring is monitored and calibrated. By 2008, however, the CLA plans to use computer software to score the performance tasks as well (see table 1).

Through its use of statistical sampling, the CLA has been able to move away from testing all students on all tasks. Unlike the twelve-hour, 3,200-item assessment used in the 1928 Pennsylvania Study, which focused on individual student development, the CLA focuses on program improvement. Accordingly, less information is provided to students confidentially
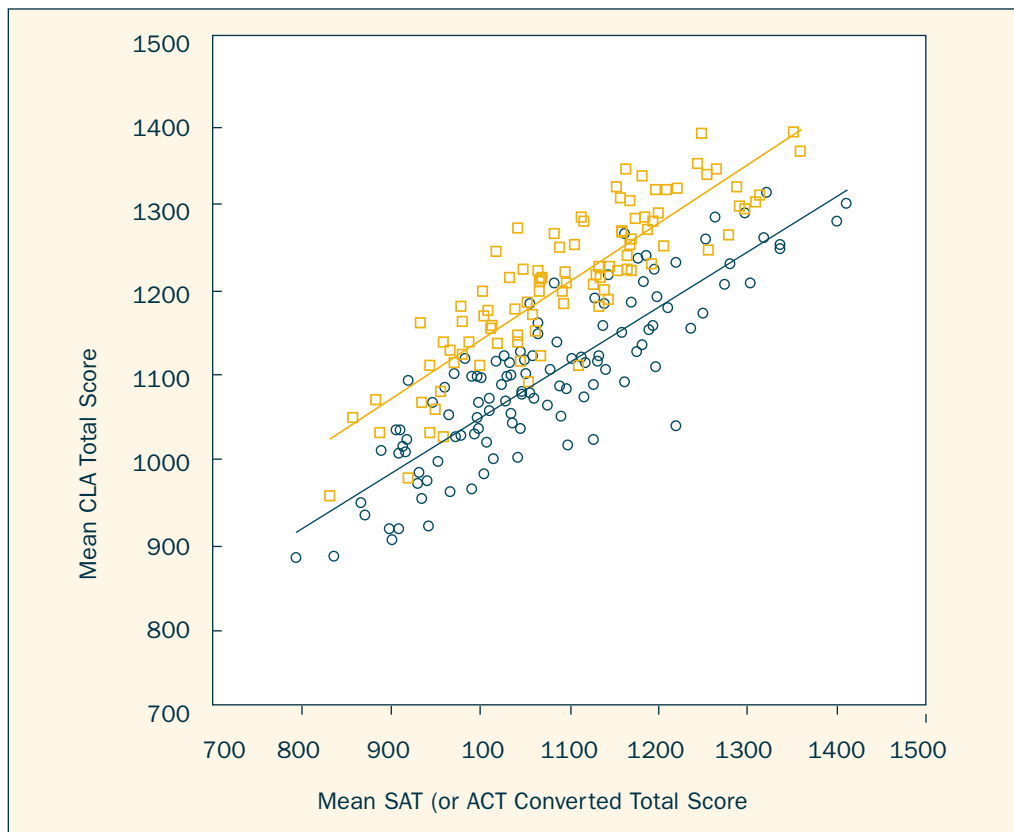

Figure 3. Relationship between mean SAT/ACT scores (in SAT units) and CLA scores [freshmen (blue, N=117 colleges) and seniors (red, N=98)]. Reproduced by permission from Council for Aid to Education, Collegiate Learning Assessment Institutional Report. (New York: Council for Aid to Education, 2006), 11.

(i.e., not available to the institution). Through cross-sectional comparisons, longitudinal cohort studies, or some combination of both, institutional (and subdivisional) reports provide a number of indicators for interpreting performance, including anonymous benchmark comparisons, percentages scoring below a certain level, and value added over and above the performance expected based on admitted-student abilities (see fig. 3).

Figure 3 shows the performance of entering freshmen (fall 2005) and seniors (spring 2006) at a set of colleges. Each point on the graph represents the average (mean) college performance on the ACT/SAT and the CLA. A number of features in figure 3 are noteworthy. First, perhaps most encouragingly, the red dots and line (seniors) fall significantly (more than 1 standard deviation) above the blue dots and line (freshmen), which suggests that college does indeed contribute to student learning (as do other life experiences). Second, most dots fall along the straight ("regression") line of *expected performance based on ability* for both freshmen and seniors—but some fall well above, and some well below. This means that, by students' senior year, some colleges exceed expected performance compared to their peers, and some perform below expectation. So it matters not only *that* but *where* a student goes to college.[7]

The assessment is divided into three parts—analytic writing, performance tasks, and biographical information. (I focus here on the first two.) Two types of writing tasks are administered. The first invites students to make an argument for or against a particular position. For example, the prompt might be as follows: "In our time, specialists of all kinds are highly overrated. We need more generalists—people who can provide broad perspectives." Students are directed to indicate whether they agree or disagree and to explain the reasons for their positions. In a similar vein, the second type of writing task asks students to evaluate an argument (see fig. 4).

Sample CLA Analytic Writing Task: Critique an Argument

A well-respected professional journal with a readership that includes elementary school principals recently published the results of a two-year study on childhood obesity. (Obese individuals are usually considered to be those who are 20 percent above their recommended weight for height and age.) This study sampled 50 schoolchildren, ages 5-11, from Smith Elementary School. A fast food restaurant opened near the school just before the study began. After two years, students who remained in the sample group were more likely to be overweight—relative to the national average. Based on this study, the principal of Jones Elementary School decided to confront her school's obesity problem by opposing any fast food restaurant openings near her school.

The performance tasks present students with real-life problems, such as that for Dyna-Tech described above, by providing an "in-basket" of information bearing on the problem (fig. 5). Some of the information is relevant, some is not; part of the task is for the students to decide what information to use and what to ignore. Students integrate information from multiple sources to arrive at a solution, decision, or recommendation. They respond in a real-life manner by, for example, writing a memorandum to their boss analyzing the pros and cons of alternative solutions and recommending what the company should do. In scoring performance, alternative justifiable solutions to the problem and alternative solution paths are recognized and evaluated.

The CLA does not pretend to be *the* measure of collegiate learning. As the Council for Aid to Education points out, there are many outcomes of a college education, and the CLA focuses on broad critical reasoning, problem solving, and communication abilities. Moreover, with its institutional (or school/college) focus, it does not provide detailed diagnostic information about particular courses or programs (unless the sampling is done at the program level). Rather, additional institutional information is needed to diagnose problems, and campuses need to test possible solutions to those problems systematically. The CLA, then, strongly encourages campuses to dig deeper.

Sample CLA Performance Task

Introductory Material: You advise Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of DynaTech's sales force, recommended that DynaTech buy a small private plane (a SwiftAir 235) that she and other members of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. Your document library contains the following materials:

1. Newspaper article about the accident
2. Federal Accident Report on in-flight breakups in single-engine planes
3. Internal Correspondence (Pat's e-mail to you and Sally's e-mail to Pat)
4. Charts relating to SwiftAir's performance characteristics
5. Excerpt from magazine article comparing SwiftAir 235 to similar planes
6. Pictures and descriptions of SwiftAir Models 180 and 235

Sample Questions: Do the available data tend to support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups? What is the basis for your conclusion? What other factors might have contributed to the accident and should be taken into account? What is your preliminary recommendation about whether or not DynaTech should buy the plane and what is the basis for this recommendation?

Reproduced by permission from Council for Aid to Education, Collegiate Learning Assessment Institutional Report. (New York: Council for Aid to Education, 2006), 7.

IV. A Proposal for Assessing Learning Responsibly

My proposal is straightforward: why not adapt the CLA framework—with its focus on broad cognitive abilities embedded in meaningful, holistic, complex tasks—and its information technologies to meet the need for assessments both within academic disciplines and of personal and social responsibility outcomes? Why not shape the writing tasks around a historical event or policy issue ripe with moral, civic, social, and personal implications? Moreover, why not set up performance tasks that provide history students an "in-basket" of information on some real or fictitious event and then ask them to review the material and resolve, as far as the evidence allows, competing interpretations of what happened? The students would be expected to present concrete evidence to support their positions and evidence to show the strengths and weaknesses of alternative positions. Why not do something similar with a civic issue, perhaps concerning the right to freedom of speech at a time when some fictitious country is engaged in a contentious war?

Admittedly, in order for this proposal to be implemented, students would have to learn to do the kinds of thinking required by the tests. This, in turn, may require changes in how students are taught—changes that may well be for the better—and careful planning to reach challenging goals. It would require that teaching, assignments, and programs be aligned with those goals so that the ambitious assessments envisioned here make sense.

For the proposal to be operational, it just might be necessary for learned societies (e.g., the American Historical Association or the Organization of American Historians), working in concert with, say, the Council for Aid to Education, to create measures that tap knowledge and reasoning within a discipline. To be sure, this would be a challenging task. Continuing with history as an example, questions of how much factual and conceptual historical knowledge should be expected of history majors in what historical domains would be hotly debated. Nevertheless, common ground could be found.

Such an assessment should be able to distinguish history majors from, say, physics majors, regardless of the particular historical topic or period. That is to say, history seniors would be expected to think like historians. They would be expected not only to take the evidence provided in the assessment "in-box," but also to recognize the context in which the event occurred and that this context probably influenced the outcome of the event. For example,

history seniors would be expected to recognize that President Benjamin Harrison's declaration that October 21, 1892, was not just the four hundredth anniversary of the discovery of America by Columbus but it was also a general holiday for the people of the United States was not just a statement about a pioneer. The declaration needs to be understood within the context of the time as both a national political statement to Italian Americans and a reflection of currents events in the world, especially in Russia.

Also at issue would be whether real historical events or fictitious events should be used. I could imagine fictitious historical events for which factually relevant and irrelevant information is provided in an "in-basket"; students would research the problem to build up an evidenced-based model to explain what happened and argue a case. Of course, some combination of actual and fictitious events might also be used.

Having learned societies engage in building assessments with an assessment development organization would create learning for both partners. It would, arguably, ensure buy-in from (some sector of) historians. It might also build the capacity of the professional organization to assist history professors in designing similar tasks that could be incorporated into their teaching and so engage their students in such learning activities. Indeed, as I talk to college faculty about learning assessment and the CLA tasks, the first thing they want to know is how to build the tasks; they believe that doing so would make an excellent teaching activity.

Similarly, the intent would not be to teach students to pass the test. Rather, the writing and performance tasks on the assessment embody the kinds of knowing and doing that faculty want their students to develop through the major. "Teaching to the test" would, then, turn out to be positive; it would involve teaching the kinds of knowledge, understanding, and reasoning skills that constitute the outcomes of historical scholarship. Students would see similar tasks on an external learning assessment but with entirely different contents and contexts. The authenticity of the assessment tasks and their link with classroom teaching and learning would create a tightly coupled assessment system for both improving teaching and learning and reporting for external accountability purposes—that is, a valid system for assessing student learning in history.

A similar scenario could be created for assessing personal and social responsibility skills. For example, in a task involving the local environment, the student might be asked to review arguments made by local environmentalists and other community members for and against removing an old dam. In the review ("in-box"), the student would find that environmentalists want to return the land to its prior state, supporting the natural streams and rivers, flora and

fauna that once thrived there, and providing hiking, camping, and fishing recreation with personal and commercial benefits. The student also would find that the environmentalists are pitted against other community members who use the man-made lake for fishing, boating, and swimming. Moreover, the student would find that many homes, restaurants, and other commercial establishments have been built up around the lake since the dam was constructed, that the dam is used to generate the county's power, and that excess energy is sold to other counties. In addition to reviewing the various arguments, the student would be asked to outline the pros and cons of removing the county's dam and to arrive at a recommendation.

V. Concluding Comments

Today's demand for a culture of evidence of student learning appears to be new, but it turns out, as we have seen, to be very old. And it is clear that cultures of evidence do not automatically lead to educational improvement—especially when what counts as evidence does not also count as education, or only counts as a part of what is expected from a college education.

What counts as evidence, as well as the conception of higher education that underlies the evidence, has shifted from one era to another over the past hundred years. Behaviorist notions of learning and achievement—such as those in which the leadership of the Carnegie Foundation was rooted and those reflected in the MAPP, CAAP, and CBASE—have competed with progressive notions—such as those embedded in the University of Chicago's general education assessment, the Cooperative Study's personal and social development tests, the Tasks in Critical Thinking, and now the CLA. No doubt these notions will continue to be contested over the course of the next hundred years.

The history of learning assessment provides some important lessons for today: Develop and justify a conceptual framework for college outcomes and for assessing achievement and learning. Design assessment systems to collect snapshots of performance both at a single point in time (achievement) and over time (learning). Tap outcomes at multiple levels, and include the difficult to assess personal and social perspectives that are so highly valued as higher education outcomes. Do this assessing in a coherent system that measures the broad spectrum of outcomes, for, if this is not done, the system's output measures will become the narrow outcomes of higher education. Finally, recognize that any large-scale assessment system can, at best, serve a signaling function; it can signal where a problem may exist, but it will not be adequate contextually and diagnostically to pinpoint the problem and generate conjectures as to how to solve it. To pinpoint problem areas, a campus needs to recognize what the external assessment can and cannot do, and it needs to have its own assessments and culture of evidence. To bring assessment information to bear on the improvement of teaching and learning broadly conceived, campus assessments need to be linked to the broader assessment system, on the one hand, and to institutional structures and processes, on the other.

On the face of it, my proposal for assessing learning could be read as audacious. However, it really is not. Rather, it is a logical step, rooted in the work of giants of the past, in assessment married to breakthrough possibilities provided by the Internet and natural language processing.

Appendix

Summary of Tests and Testing Programs by Era

Era	Study Program, or Test Provider	Test
The Origins of Standardized Tests of Learning: 1900–1933	Missouri Experimental School Study	Objective Tests of arithmetic, spelling, penmanship, reading, English, and composition
	Thorndike MIT Engineers Study	Objective tests of mathematics, English, and physics
	Pennsylvania Study	Objective tests of general culture (literature, fine arts, history and social studies, general science), English (e.g., spelling, grammar, vocabulary), mathematics, and intelligence
The Assessment of Learning in General and Graduate Education: 1933–47	Chicago College General Education	Constructed response and objective tests focused on analysis, interpretation, and synthesis
	Cooperative Study of General Education	Objective tests of general culture, mathematics, and English (based on the Pennsylvania Study); inventories of general life goals, satisfaction in reading fiction, social understanding, and health
	Graduate Record Examination (GRE) Program	1936: Objective Profile Tests of content (e.g., mathematics, physical sciences, social studies, literature, and fine arts) and verbal ability (see Pennsylvania Study)
		1939: Above plus sixteen Advanced Tests in major fields (e.g., biology, economics, French, philosophy, sociology) for academic majors
		1946: General Education Tests that included the Profile Tests plus "effectiveness of expression" and a "general education index"
		1949: Verbal and Quantitative Aptitude Tests, created as stand-alone tests, replaced the Verbal Factor Test and the Mathematics Test in the Profile Tests
		1954: Area Tests were "entirely new measures of unusual scope . . . [providing] a comprehensive appraisal of the college student's orientation in three principal areas of human culture: social science, humanities, and natural science" (ETS 1954, 3). These three new tests replaced the Profile and General Education Tests

Era	Study Program, or Test Provider	Test
The Rise of the Test Providers: 1948–78	ETS	Undergraduate Assessment Program that include the GRE tests
	ACT	College Outcomes Measures Project from constructed response tests to objective tests to save time and cost
	New Jersey	Tasks in Critical Thinking constructed response tests
The Era of External Accountability: 1979–Present	ETS	Academic Profile and Measure of Academic Proficiency and Progress (MAPP) largely objective tests
	ACT	College Assessment of Academic Proficiency (CAAP) largely objective tests
	CAE	Collegiate Learning Assessment constructed response tests

Notes

1. While Pritchett was skeptical of the capacity of objective tests to predict success in college, the intent to shift from "discursive examinations" or essays to "objective testing" was clear in his observation that "there are few devices connected with teaching more unsatisfactory than our present day examinations, whether used as tests for admission or as criteria of performance on the part of the student" (Savage 1953, 285). The Carnegie Foundation for the Advancement of Teaching was committed.

2. "Assessment of learning" is a catch phrase that includes indirect and direct measures of learning. The phrase is understood by the public and policy makers; hence, it communicates its intent to focus on important outcomes—student learning—and not simply on inputs and processes. However, this phrase is technically incorrect. Learning is a relatively permanent change in a student's behavior over time. Typically, achievement is assessed—the accumulation or amount of learning up to the point in time when performance is measured.

3. While the Pennsylvania Study was underway, the Progressive Education Association (PEA) launched its Eight Year Study (1930). The PEA study aimed to reform high school education to meet the needs of what had become a disengaged student body. Both Learned and Wood served as members of the commission for the study, but not as members of the directing subcommittee. The study took a different approach to assessing student learning than did the Pennsylvania Study. It included constructed-response test items as well as objective test items, and it went beyond academic skills to include life-skills in both the cognitive and personal and social responsibility domains. The PEA and its Eight Year Study, then, challenged the very ideas espoused by the Carnegie Foundation. This difference in views—focus on declarative and procedural knowledge versus focus on broad abilities including both academic and responsibility skills—persists in debates about what counts and is measured as learning today.

4. The ubiquitous "Bloom's Taxonomy" grew out of the work of the examiner's office to broaden the measurement of academic achievement. Benjamin Bloom was one of the directors of the office.

5. For a short time, the Collegiate Learning Assessment (CLA) used some tasks from the Tasks in Critical Thinking. These tasks have now been replaced by a full set of CLA-developed performance tasks.

6. The Council for Aid to Education (CAE), in developing the Collegiate Learning Assessment (CLA), also borrowed from the constructed-response essay questions recently added to the Graduate Record Examination (GRE) when the analytic reasoning multiple-choice test was dropped. Today, GRE prompts have been replaced in the CLA by prompts and scoring developed by CAE.

7. Technically, there are multiple possible explanations that challenge the college-effect interpretation. Some selection must go on between the freshman and senior years with dropouts and transfers that are not perfectly picked up by the ACT/SAT test. Or simply maturing in the everyday world would provide those skills. Or college is not the only societal mechanism for performance improvement; the same young adults would have developed these very same capabilities on the job or in the military. Nevertheless, we find the college-effect interpretation the most plausible, but not the only possible, explanation.

References

American Association of State Colleges and Universities (AASCU). 2006. Value-added assessment: Accountability's new frontier. *Perspectives* (Spring): 1–16.

Association of American Colleges and Universities (AAC&U). 2005. *Liberal education outcomes: A preliminary report on student achievement in college.* Washington, DC: Association of American Colleges and Universities.

———. 2007. *College learning for the new global century: A report from the National Leadership Council for Liberal Education and America's Promise.* Washington, DC: Association of American Colleges and Universities.

Baxter, G. P., and R. J. Shavelson. 1994. Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research* 21 (3): 279–98.

Dunkel, H. B. 1947. *General education in the humanities.* Washington, DC: American Council on Education.

Dwyer, C. A., C. M. Millett, and D. G. Payne. 2006. *A culture of evidence: Postsecondary assessment and learning outcomes.* Princeton, NJ: Educational Testing Service.

Educational Testing Service (ETS). 1953. *The Graduate Record Examination Institutional Testing Program: Summary statistics.* Princeton, NJ: Educational Testing Service.

———. 1966. *The Graduate Record Examinations: Area tests.* Princeton, NJ: Educational Testing Service.

———. 1994. *Tasks in critical thinking: Using an innovative extended task format for performance-based assessment of critical thinking skills.* Princeton, NJ: Educational Testing Service.

Erwin, T. D., and K. W. Sebrell. 2003. Assessment of critical thinking: ETS's Tasks in Critical Thinking. *The Journal of General Education* 52 (1): 50–70.

Ewell, P. 2001. Statewide testing in higher education. *Change* 33 (2): 21–27.

Executive Committee of the Cooperative Study in General Education. 1947. *Cooperation in general education.* Washington, DC: American Council on Education.

Frodin, R. 1950. Very simple but thoroughgoing. In *The idea and practice of general education: An account of the College of the University of Chicago by present and former members of the faculty*, ed. F. Ward, 25–99. Chicago: The University of Chicago Press.

Immerwahr, J. 2000. *Great expectations: How New Yorkers view higher education.* San Jose, CA: The National Center for Public Policy and Higher Education.

Kandel, I. L. 1936. *Examinations and their substitutes in the United States.* Bulletin Number Twenty-Eight. New York: The Carnegie Foundation for the Advancement of Teaching.

Lagemann, E. C. 1983. *Private power for the public good: a history of the Carnegie Foundation for the Advancement of Teaching.* Middleton, CT: Wesleyan University Press.

Lannholm, G. V., and W. B. Schrader. 1951. *Predicting graduate school success: An evaluation of the effectiveness of the Graduate Record Examinations.* Princeton, NJ: Educational Testing Service.

Learned, W. S., and B. D. Wood. 1938. *The student and his knowledge: A report to the Carnegie Foundation on the results of the high school and college examination of 1928, 1930, and 1932.* No. 29. Boston: The Merrymount Press.

Levi, A. W. 1948. *General education in the social studies.* Washington, DC: American Council on Education.

Lythcott, J. 1990. Problem solving and requisite knowledge of chemistry. *Journal of Chemical Education* 67: 248–52.

Martinez, M. E. 1999. Cognition and the question of test item format. *Educational Psychologist* 34: 207–18.

National Research Council. 2001. *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Ruiz-Primo, M. A., R. J. Shavelson, M. Li, and S. E. Schultz. 2001. On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment* 7 (2): 99–141.

Savage, H. J. 1953. *Fruit of an impulse: Forty-five years of the Carnegie Foundation.* New York: Harcourt, Brace, and Company.

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

———. 2006. A brief history of assessing undergraduates' learning: Carnegie Foundation for the Advancement of Teaching's heritage. Paper presented at the Carnegie Centennial Conference on Improving Quality and Equity in Education, Princeton, NJ.

Warren, J. R. 1978. *The measurement of academic competence.* Princeton, NJ: Educational Testing Service.

About the Author

Richard J. Shavelson is the Margaret Jack Professor of Education, professor of psychology, and senior fellow in the Woods Institute for the Environment at Stanford University. His current work includes the assessment of undergraduates' learning, the assessment of science achievement, and the study of inquiry-based science teaching and its impact on students' knowledge structures and performance. Other work includes studies of computer cognitive training on working memory, fluid intelligence and science achievement, accountability in higher education, scientific basis of education research, and new standards for measuring students' science achievement in the National Assessment of Educational Progress (the nation's "report card"). His publications include *Statistical Reasoning for the Behavioral Sciences, Generalizability Theory: A Primer* (with Noreen Webb) and *Scientific Research in Education* (edited with Lisa Towne); he is currently working on a book tentatively titled *The Quest to Assess Learning and Hold Higher Education Accountable*. He is a fellow of the American Association for the Advancement of Science, the American Psychological Association, the Association for Psychological Science, and the Humboldt Society.

AAC&U is the leading national association concerned with the quality, vitality, and public standing of undergraduate liberal education. Its members are committed to extending the advantages of a liberal education to all students, regardless of academic specialization or intended career. Founded in 1915, AAC&U now comprises more than 1,100 accredited public and private colleges and universities of every type and size. AAC&U functions as a catalyst and facilitator, forging links among presidents, administrators, and faculty members who are engaged in institutional and curricular planning. Its mission is to reinforce the collective commitment to liberal education at both the national and local levels and to help individual institutions keep the quality of student learning at the core of their work as they evolve to meet new economic and social challenges.

Information about AAC&U membership, programs, and publications can be found at www.aacu.org.

Association of American Colleges and Universities