Group of National Experts on the AHELO Feasibility Study

THE COLLEGIATE LEARNING ASSESSMENT (CLA) INITIATIVE IN THE UNITED STATES

Prepared by Roger Benjamin (Council for Aid to Education) and Richard Shavelson (Stanford University)

17-18 December 2008

The AHELO GNE is invited to TAKE NOTE of the CLA initiative.

This document is available only in PDF format.

Cancel and replace: to include the annexes.

Contact: Karine Tremblay, Directorate for Education
[Tel: +33 1 45 24 91 82; Email: Karine.Tremblay@oecd.org]
THE COLLEGIATE LEARNING ASSESSMENT (CLA) INITIATIVE IN THE UNITED STATES

1. This set of papers provides an overview of the background, history and characteristics of the Collegiate Learning Assessment initiative in the United States.

2. The AHELO GNE is invited to TAKE NOTE of the CLA initiative.
Over the past 35 years, the public, as well as state and federal policy makers, have increasingly pressured higher education to create a culture of evidence and, specifically, to account for student learning. While virtually all states report on collegiate learning using proxies (e.g., graduation rates), colleges and universities are now being asked to assess 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’ capacity to think critically, solve problems, and communicate analytically and clearly.
The commission has opened a conversation that we in higher education need to continue—a conversation about what our students should be learning and what the history of measurement tells us about how we should (and shouldn’t) determine whether they have learned. Here I would like to suggest what some of those goals might be, review that history, and end by suggesting how we might approach the task of assessing the full range of collegiate learning.

**LEARNING OUTCOMES**

There is more to be learned in college than the broad abilities singled out by the Secretary’s commission. The full range of outcomes for college students goes from factual, procedural, and conceptual knowledge and reasoning in a discipline, to thinking that might be applied to a very wide variety of situations, to the development of intelligence. Moreover, “cognitive” outcomes include “soft skills,” in that reasoning is involved in personal relations, moral challenges, and civic engagement. Chart 1 suggests the interrelationships among these kinds of learning.
If the learning outcomes of higher education are narrowly measured, as cost, capacity, and convenience would dictate, we risk narrowing the missions, subject matter taught, and diversity of the American system of higher education. For example, it is unsatisfactory to say that we should not measure soft skills 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, & Payne, 2006, p. 20). If we do not assess them, these crucial skills will drop from sight as pressures for accountability lead campuses to focus on a more restricted, more easily and less expensively measured subset of learning outputs, such as multiple-choice tests of declarative content knowledge. What we need to do is to learn from the history of assessing learning and take audacious steps to develop and measure the full set of learning outcomes that our nation’s public so highly values.

This demand to establish a culture of evidence has a long lineage, and the future of this culture may depend on how well we understand the past. The key lesson we learn from that history is this: Cultures of evidence do not automatically lead to educational improvement if what counts as evidence does not also count as education or counts as only part of it.

**A Brief History of Learning Assessment**

Spurred by the success of standardized “objective” mental testing in World War I (namely the Army Alpha test with multiple-choice and true-false questions having one correct answer), the first third of the 20th century marked the beginning of testing learning in higher education. During this period, the emphasis was on the mastery of academic content.

The Carnegie Foundation for the Advancement of Teaching led the movement. Howard Savage, a staff member and historian of the foundation in its early days, attributes Carnegie’s leadership in assessing college students’ learning 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:

> [T]he 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 indi-
knowledge and abilities than the Penn nations tested a much broader range of knowledge was feasible. The most note- comprehensive assessment of content-cepts, and mathematical routines.

Education program. The Chicago exami- the University of Cincinnati, and Co- Massachusetts Institute of Technology,-lumbia University concerning facts, con-cept, and mathematical routines.

The Pennsylvania study proved that comprehensive assessment of content knowledge was feasible. The most noteworthy next step taken in testing during the progressive period between 1933 and 1947 was the attempt to assess not only cognitive outcomes beyond factual and procedural knowledge, but also the personal, social, and moral outcomes of general education.

During this period, general education and general colleges, such as the University of Chicago’s progressive College, sprouted in universities across the country. In the Chicago program, a central University Examiner’s Office, rather than the faculty, was responsible for developing, administering, and scoring tests in the university’s general-education program. The Chicago exami-nations tested a much broader range of knowledge and abilities than the Penn-sylvania study had: the use of knowledge in a variety of unfamiliar situations; the ability to apply principles to explain phenomena; and the ability to predict outcomes, determine courses of action, and interpret works of art. Open-ended essays and multiple-choice questions demanding interpretation, synthesis, and application of new texts (primary sourc-es) characterized the comprehensive exams. The ubiquitous “Bloom’s Taxonomy” of cognitive objectives (knowledge, comprehension, application, analysis, synthesis, evaluation) grew out of the work of the Examiner’s Office.

Meanwhile, the Cooperative Study of General Education, conducted by a consortium of higher-education institutions, reflected progressive notions of human development as well. Members of the consortium believed they could benefit from a cooperative attempt to improve general education. To this end they assessed 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.

But even as these expanded no- tions of learning were being acted on, Learned parlayed his experience with the Pennsylvania study into an assessment for graduate education. In proposing the “Cooperative Graduate Testing Program,” Learned noted that with increased demand for graduate educa- tion following the Depression and with the decreasing credibility of the bac-calaureate degree, something more than the number of college credits a student had earned was needed on which to base graduate admissions decisions.

In concert with the graduate schools at Columbia, Harvard, Princeton, and Yale, in 1937 Learned’s team adminis- tered seven tests—the first administra- tion of what was to be the Graduate Record Examination (GRE). From there the program grew by leaps and bounds (Chart 2, p. 30). Despite its success, the growing financial and logistical burdens it imposed led the Carnegie Foundation to spin off an independent national testing agency, the Education Testing Ser- vice (ETS), to administer the exam.

The original GRE, like the Penn-sylvania study’s examination, was a comprehensive, objective test focused largely on students’ content knowledge, but it also tapped verbal reasoning and was used to infer students’ fitness for graduate study. In 1949 the GRE also began to test general reasoning via the GRE Aptitude Test, with the verbal and quantitative sections we see today.

ETS continued the shift away from content toward general reasoning in 1954 by replacing the profile tests and
the tests of general education with “area tests” that focused on the social and natural sciences and humanities, as a means of assessing the broad outcomes of the liberal arts. The tests emphasized reading comprehension, understanding, and interpretation, often providing the 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, p. 3).

When the Carnegie Foundation moved the GRE to ETS, it left an extraordinarily strong legacy that has endured into the 21st century: objective, group-administered, cost-efficient testing using selected response—now solely multiple-choice—questions. Testing organizations developed precursors to multiple-choice questions. Testing using selected response—now solely objective testing—was introduced in the 1960s and 1970s, including ETS’s Undergraduate Assessment Program, which incorporated the GRE, and American College Testing’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 into today’s Collegiate Assessment of Academic Proficiency (CAAP)—multiple-choice test batteries designed to measure and improve general-education outcomes.

However, several legacies of the progressive era emerged in the late 1970s, augmenting a change in the course set by Learned. 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. This led to several developments. ETS studied “constructed-response” tests that tapped communication skills, analytic thinking, synthesizing ability, and social/cultural awareness, while American College Testing (ACT) tried out open-ended, performance-based assessments of skills for effective functioning in adult life. And New Jersey developed Tasks in Critical Thinking, which asked students to perform real-world tasks in a “performance-based assessment ... [of] the skills of inquiry, analysis, and communication” (ETS, 1994, p. 2), with prompts that did “not assess content or recall knowledge” (p. 2).

For a short period of time, these assessment programs set the mold. But due to time limitations, as well as issues with scoring, reliability, and cost, they either faded or morphed back into multiple-choice tests. By the beginning of the 1980s, political pressure to assess student learning and hold campuses accountable for it led to an increased demand for standardized, multiple-choice testing (e.g., in Florida and Tennessee).

By the mid-80s, though, some states (e.g., Virginia, South Carolina, and Missouri) had mandated not standardized testing but campus-based assessment, to which the campuses responded by creating individualized assessment programs. Some of those programs relied on home-grown assessments, while others looked to the testing companies to provide measures. By this time, a wide array of college-learning assessments in the Carnegie tradition was available, which remain popular to this day.

In the area of general education, ETS currently provides the MAPP and ACT administers the CAAP. The College Resource Center at the University of Missouri, Columbia, offers the College Basic Academic Subjects Examination (CBASE), a criterion-referenced achievement examination in English, mathematics, science, and social studies that serves to qualify individuals for entry into teacher-education programs and as a test of general academic knowledge and skills.

But as we saw at the end of the 1970s, objective testing was not the way faculty members wanted student learning to be assessed. They were more comfortable with open-ended, holistic, problem-based assessments, which were more in tune with what they thought they were teaching. Intuitively, they suspected that the kind of thinking that is stimulated and the type of performance that is assessed by multiple-choice and other highly structured tests differ from that generated 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 reasoning is involved. Students’ performance varies considerably when the same task is presented as a multiple-choice question, an open-ended question, or a concrete performance task.

The current (November 2006) draft report of the Secretary of Education’s higher-education commission specifically mentioned two tests as examples when it recommended that “postsecondary education institutions should measure and report meaningful student learning outcomes”: the MAPP and the Collegiate Learning Assessment (CLA). In supporting the commission’s recommendation, the American Association of State Colleges and Universities zeroed in on the CLA, giving a rationale for that choice that echoes the progressive-era arguments that are consonant with faculty preferences and with best practices in teaching and learning.

**The Collegiate Learning Assessment**

The Council for Aid to Education developed the Collegiate Learning

---

**Chart 2. GRE’s Growth over Its First Decade**

![Chart showing growth of GRE over its first decade](image)

*Source: Richard J. Shavelson*
Assessment (CLA). As the CLA may provide insight into a next generation of learning assessments, I note its major features here (see table, p. 32).

Just as objective testing and computer scoring of response sheets (developed by IBM for Learned) revolutionized assessment of learning at the turn of the 20th century, so too has new information technology and statistical sampling ushered in a significant change in assessing college students’ learning at the turn of the 21st century. The CLA moves away from a selected-response, multiple-choice strategy to one that test developers from the progressive era would have recognized—asking students to perform concrete, complex, open-ended tasks.

The CLA’s notions of learning, as well as of assessment, had their origin in that same period. The CLA assesses not content knowledge but the skills of critical thinking, analytic reasoning, problem solving, and written communication. These capabilities are tapped in realistic “work-sample” tasks drawn from education, work, and everyday life. They are accessible to students from a wide diversity of majors and general-education programs.

The assessment is divided into two parts (plus biographical information): analytic writing and performance tasks (for examples of both, go to http://www.cae.org/content/pdf/CLA.in.Context.pdf). Two types of writing tasks are administered. The first, “make an argument,” invites students to present an argument for or against a particular position. The second, “break an argument,” asks students to critically evaluate an argument.

The performance tasks present real-life problems to students, along with resources to solve them, such as an “in-basket” (or nowadays, “computer basket”) with packets of information, some of which bear on the problem and some of which are irrelevant. (Part of the assessment is for the students to decide what information to use and what to ignore.) Students integrate these multiple sources of information to arrive at a solution, decision, or recommendation. Students respond in realistic ways—for example, by writing a memorandum analyzing the pros and cons of alternative solutions, and recommending what a hypothetical company should do. In scoring performance, alternative justifiable solutions to the problem and alternative paths to solving it are recognized and evaluated.

The test can be scored reliably without incurring the prohibitive costs of traditional essay testing due to recent developments in information technology. 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. Performance tasks, currently scored by human raters, will also be scored by computer software within a year.

Sampling is another key feature that makes CLA-type tests feasible. Students are no longer willing to sit for 12-hour exams, as they once did in Pennsylvania. The capacity to provide these rich tasks without overburdening students or increasing costs to unsustainable levels is a function of statistical sampling (all students need not be tested on all tasks), which is made possible by the different goals of the testing. The basic focus in Pennsylvania was on individual students’ development; CLA focuses on program improvement.

Institutional (and subdivision) reports provide a number of indicators for interpreting performance, including anonymous benchmark institutional comparisons and the percentage of institutions scoring below a certain level. The CLA also enables institutions to determine the value they add to students’ learning, over and above the performance expected at the institution based on the abilities of admitted students, cross-sectional comparisons, longitudinal cohort studies, or some combination.

Chart 3 (p. 33) shows the performance of a set of colleges’ freshmen and seniors in spring 2005. Each point on the graph represents average (mean) college performance on the ACT/SAT and the CLA; the green dots and line are seniors, and the purple dots and line are freshmen.

A number of features in Chart 3 are noteworthy. First, perhaps most encouragingly, the green dots and line (seniors) fall significantly (more than 1.6 standard deviations) above the purple 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 the senior year some colleges exceed expected performance compared to their peers, and some perform below expectation. So it matters not only that you go, but also where you go to college.

The CLA does not pretend to be the measure of collegiate learning. Rather, as the Council for Aid to Education points out, the CLA focuses on the broad abilities of critical reasoning, problem solving, and communication, which do not exhaust the list of collegiate learning outcomes. Moreover, with its institutional (or school/college) focus, it does not provide detailed, diagnostic information about particular courses or programs (unless sampling is done at a program level). Other institutional information is needed to diagnose problems, and campuses need to test possible solutions to those problems systematically. The CLA, then, sends a strong signal to the campus to dig deeper.

**An Audacious Proposal for Assessing Learning Responsibly**

Why not use a CLA-type framework—with its focus on broad cognitive abilities embedded in meaningful, holistic, complex tasks and its use of information technologies—to assess academic programs? That is, why not shape a writing task around, for example, an historical event? History students could be given a computer basket of information on the event and asked to adjudicate among competing interpretations of what happened. It might be possible for disciplinary societies, in this case the American Historical Association, to work with the Council for Aid to Education to create measures that tap knowledge and reasoning within the discipline.

To be sure, this would be a challenging task. Issues of how much factual and conceptual historical knowledge, and in what domains of history to require for successful completion of the task would be hotly debated. But the writing and performance tasks would not be designed to elicit students’ declarative content knowledge so much as to embody the kinds of thinking that faculty want their students to develop within the major. For that purpose, both real and fictitious events, or a combination, could be used. Having learned societies engage in building assessments with an assessment-development organization would create learning for both partners, create some disciplinary buy-in, and build the capacity of the society to assist professors in designing similar tasks for their courses.

### Characteristics of the Collegiate Learning Assessment

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Attributes</th>
</tr>
</thead>
</table>
| Open-ended Tasks     | • Critical thinking, analytic reasoning, problem-solving and written communication are assessed  
                       • Realistic work samples  
                       • Engaging tasks  
                       • Applicable to different academic majors |
| Computer Technology  | • Interactive Internet platform  
                       • Paperless administration  
                       • Natural-language processing software for scoring writing  
                       • Human scoring of performance tasks, currently  
                       • Report of institution’s (and subdivision’s) performance (and individual student performance confidentially to student) |
| Focus                | • Institutional or school/department/program  
                       • Not on individual student performance |
| Sampling             | • Not all students perform all tasks  
                       • Creates scores at institution or subdivision/program level as desired (depending on sample sizes) |
| Reporting            | • Controls for student ability so that “similarly situated” benchmark campuses can be compared  
                       • Provides value-added estimates  
                       • Provides percentiles  
                       • Provides benchmark institutions |

Source: Richard J. Shavelson
CLA-type measures could also be used to assess what the AAC&U calls individual and social-responsibility outcomes. Indeed, some of the tasks on the current CLA tap some of these competencies. Why not create more such tasks—for instance, one that focuses on a civic issue, perhaps regarding the right to freedom of speech at a time when a fictitious country is engaged in a contentious war or about the effectiveness of a group of local environmentalists’ argument for removing an old dam?

**CONCLUSION**

Today’s demand to establish a culture of evidence about student learning appears to be new, but it is not. What is clear from this review of past and current attempts to assess college students’ learning is that cultures of evidence do not lead to educational improvement if what counts as evidence does not count as education or only counts as a part of what we expect from a college education.

Over the past 100 years, two notions about what college students should learn and how they should be tested have competed: the behaviorist notions of achievement reflected in Carnegie’s leadership and embedded in the MAPP, CAAP, and CBASE have vied with the progressive notions of learning that we see in the University of Chicago’s general-education assessment, the Cooperative Study’s personal and social-development tests, and attempts to assess general intellectual skills such as the Tasks in Critical Thinking and now the CLA. No doubt these views will continue to tussle for the next century. That said, my bet is on CLA-type assessments.

The history of assessing learning provides some important lessons. It is important that we develop and justify a conceptual framework for college outcomes and assessment of achievement and learning. We must design assessment systems to collect both snapshots of performance at one point in time (achievement) and over time (learning). We need to include in those assessments the broad spectrum of collegiate outcomes, including the difficult-to-measure personal and social perspectives that we value so highly, since what we measure is what we will come to teach.

Finally, we need to recognize that any large-scale assessment system can, at best, signal where a problem may exist, but it will not pinpoint the problem and generate solutions. For this, a campus needs to place any external assessment in the context of its own rich array of assessments and link it to institutional structures and processes. Only by doing this will it be able to bring assessment information to bear on the improvement of teaching and learning.

---

**Chart 3. Relationship Between Mean SAT (and ACT Converted) and CLA Scores**

![Chart showing the relationship between Mean SAT score and Mean CLA Total Score for Freshmen and Seniors.](image)

**Source:** Council for Aid to Education

---

**Resources**

The Case for Comparative Institutional Assessment of Higher-Order Thinking Skills

By Roger Benjamin

Times of threshold change, such as the transformation from the industrial era to the knowledge economy of today, produce pressures to redesign the institutions we live with to respond to or, better, shape this change. In America’s knowledge economy, there is broad agreement that the only way to preserve the nation’s economic edge will be through constant innovation, the creation of ideas that produce new economic value, for which a highly educated workforce is necessary. So, since education is the principal venue for human capital development, it is not surprising that the public gaze has turned to schools and colleges. Specifically, the public and its representatives want to know about the nature and quality of educational outcomes that those institutions generate.

In this context, one of the most contentious questions currently being debated is whether or not it is possible or desirable to assess and publicly communicate about learning outcomes in a way that permits comparisons between institutions. I contend that comparative assessments are essential and practical: although we cannot measure all dimensions of learning, it is perfectly possible to comparatively assess the higher-order thinking skills that are etched in most collegiate mission statements, thought to be particularly important in the knowledge economy, and shared by most educators as key aims of instruction: critical thinking, analytical reasoning, problem solving, and writing. I also argue that such assessments are a critical part of larger efforts to develop better approaches to teaching and learning—goals all educators share. And finally, I make the case for assessing learning in a way that reveals the “value added” by the institution to students’ intellectual development.

Roger Benjamin is president of the Council for Aid to Education (CAE), which developed and administers the Collegiate Learning Assessment (CLA). He served previously as director of RAND Education, provost at the Universities of Minnesota and Pittsburgh, and professor of political science at the University of Minnesota.
WHY COMPARISON IS ESSENTIAL

Higher education adheres to a transparent system for the evaluation of research proposals and reports based on peer review. Reviewers selected by independent funding agencies such as the National Science Foundation and the National Institutes of Health make decisions about research funding, and peer editors select manuscripts for publication, on a competitive basis.

The consequences include rankings of the standing of colleges and universities based on the amount of federal money, publications, citations, and awards that faculty obtain. While these systems to rank institutions' research preeminence are far from ideal, few would argue that the quality of the American research enterprise does not benefit from this system of judgment based on the ability to compare.

Contrast the situation in research to that in undergraduate education. Until the recent development of the Voluntary System of Accountability (VSA) [Editor's note: see the interview of George Mehaffy and David Shulenberger on the VSA in the July/August 2008 issue of Change] by the National Association of State Universities and Land Grant Colleges (NASULGC) and the American Association of State Colleges and Universities (AASCU), higher education leaders resisted the idea that comparative information about student learning should be collected for public consumption.

John Lombardi, for example, argued in Inside Higher Education (August 2, 2006) that we should not benchmark because we already have a higher education market in which students are free to apply to the schools of their choice. In Lombardi's view, the fact that students of high ability apply to institutions that charge the highest tuition provides us with enough information about the quality of the education they will receive there.

But the effectiveness of market measures to ensure quality requires that the consumer be provided with useful information about the product compared to other choices he or she might make. As Peter McPherson and David Shulenburger, president and vice president of NASULGC respectively, argue (2006), comparative public information about not only student learning outcomes but class size, student services, per-student endowment, and graduation rates will give consumers (or at least the less than 20 percent of prospective students who have a choice in the college they attend) the means to make better decisions.

Between-institution comparison is particularly critical now because of the formative stage in which we are, with respect to the development of useful theories of pedagogy and learning. The university is an organization engaged in continuous improvement in research, a byproduct of the scientific method and rules of scholarly inquiry. It is difficult, however, to find colleges and universities that are engaged in the continuous improvement of teaching and learning. We do not have agreed-upon methods of teaching, strategies for assessment, or standards of learning that we think students should be held accountable to. Under these conditions, comparative researchers such as Charles Ragin and others argue that it is critical to expand the universe of cases that might be compared with each other. From this larger universe of cases the researcher can identify outliers that warrant greater examination as possible best practices.

One argument against comparing institutions rests on Ernest Pascarella’s, Patrick Terenzini’s, and George Kuh’s research, which suggests that within-institution variation with respect to student learning and engagement is greater than between-institution variation. As Pascarella and Terenzini write:
The weight of evidence from the 1990s casts considerable doubt on the premise that the substantial structural, resource, and qualitative differences among postsecondary institutions produce correspondingly large differences in net educational effects on students. Rather, the great majority of postsecondary institutions appear to have surprisingly similar net impacts on student growth. ... Similarities in between-college effects substantially outweigh the differences (p. 590).

If colleges and universities differ less than does student experience within any one of them, the argument goes, one should focus only on internal comparisons.

But Pascarella and Terenzini’s review of the learning literature ends at the conclusion of the 1990s. They did not have the available evidence from the comparative measurement instrument with which I am most familiar, the Collegiate Learning Assessment (CLA). The CLA shows not only substantial growth within institutions but differential growth among them. While there is over 1.0 standard deviation value-added growth in CLA scores within institutions (estimated cross-sectionally), there are up to three standard deviation differences in the results between institutions with similar SAT scores—a large effect, and much greater than within-institution growth. It would also be useful to see how other comparative measures of student learning such as the Measure of Academic Proficiency and Progress (MAPP) and the College Assessment of Academic Progress (CAAP) distinguish among institutions.

This shows that it does matter where one goes to college. There are significant differences among institutions, at least with respect to their ability to improve higher-order skills. Again, the variance suggests that there are interesting best practices to be studied in those that are doing better than expected, practices that can be adapted by all colleges focused on the improvement of teaching and learning.

One attribute usually and correctly cited in assertions about the superiority of American higher education is its diversity, which is also used to make the case that colleges and universities should not be compared (in contrast to the argument based on the assertion that they are not different enough). Certainly, collegiate missions range widely: some are public and others private; some are nonprofit and others for-profit; some have a liberal arts orientation, while others focus on science and technology; business, music, and the other performing arts, fine art, architecture, and so on. These distinct missions clearly result in different emphases—configurations of majors and minors, types of core curricula, student populations served, and expectations of student performance, among others. The implication of this description of American higher education suggests to some that any effort to compare such a large group of institutions would be fatally reductionist. But to me it suggests that if one argues for comparative assessment, a clear corollary is that the focus needs to be on a learning goal that virtually all institutions share—such as the development of higher-order skills.

**WHY HIGHER-ORDER SKILLS SHOULD BE THE PRINCIPAL FOCUS, AND THE INSTITUTION THE INITIAL UNIT OF ANALYSIS**

The academic disciplines were clearly the most important venues for learning during the industrial era, when the transfer of knowledge (content) was thought to be the primary function of undergraduate education, and disciplinary courses have long been the basic building blocks of the curriculum (although majors account for less than one-third of the courses students take in a typical four-year program). But while students still need to acquire a body of knowledge, it is equally, if not more, important for them to master the higher-order skills they will need to access, structure, and use information (James Pellegrino, et al.). According to John Immerwahr, employers are not as concerned about what students major in as they are about how well and flexibly they think. Under our present organizational model, faculty have no way of knowing whether their courses—delivered within departments and isolated from one another—in fact develop these greatly valued skills in their students.

The institution as a whole is the appropriate level at which to evaluate success in improving higher-order skills in its students, because it is at the level of the mission statement that virtually all institutions commit to improving these higher-order skills. This does not mean the majors are not important in contributing to students’ intellectual development; they are crucial to students’ mastery of both disciplinary skills and content. Moreover, it is important to drill down to the individual classroom to understand what changes in pedagogical practice should be effected in order to improve student learning. We need to identify the systematic steps that link the institution-level results to the faculty in their classrooms in order to complete a continuous system of improvement of teaching and learning. But students develop in a cumulative fashion as they progress through their courses and other experiences at their institutions, so the first goal is to determine the degree of improvement of those skills in the course of students’ entire baccalaureate education. This is how we can ascertain the value added by a particular college or university.

**WHY COMPARATIVE TESTS SHOULD SUPPLEMENT, NOT SUPPLANT, LOCAL ASSESSMENTS**

Unique assessments tailored to the individual needs of individual programs are often defended as the only way to make improvements in educational practice. And indeed, faculty generally respond to accreditation requirements to track student learning and other indicators of program effectiveness using a mix of locally developed indicators. Undoubtedly, the information generated by these instruments is of great value in improving programs.

The question is whether they provide adequate assistance to faculty and administrators in their efforts to improve teach-
ing and learning, or whether a measure that enabled them to compare their students' performance to that of similar students at other institutions would give them valuable additional information about their campus's effectiveness. I argue that it would and does. But I would also add that if comparative measures are to be successful, faculty must see them as relevant to their classroom work. In other words, the measures need to be constructed so as to contribute to teaching and learning—otherwise the faculty, the ultimate "customer" assessment of the information supplied, will ignore it.

Comparative measures can produce important signals to the faculty and administration of institutions about where they stand in comparison to other similar institutions. But they should not substitute for locally developed instruments. They should be only one of many an institution uses to assess its students' learning gains, since no one test can measure everything that is important or provide, by itself, enough information to guide program improvement.

And the results of comparative assessments should not be used to rank or rate colleges and universities. Single-dimension rankings mask the great diversity of inputs and outcomes in higher education. Moreover, the field of assessment in higher education is far from reaching maturity, and an over-reliance on single measures that are in the process of being tested and improved provides a sense of certitude where it is not warranted.

**WHY THE VALUE-ADDED APPROACH IS IMPORTANT**

An important question that every institution needs to ask itself is how much its teaching improves students' intellectual capacities. This is the value-added question that surfaces in today's assessment and accountability debate. It can do this by assessing the performance of freshmen and seniors at the same time (cross-sectional assessment) or by assessing the performance of students when they enter the institution and those same students again when they are about to graduate (longitudinal assessment). Then, in order to know whether the degree of improvement it makes in those capacities is satisfactory, given what is possible with the students it accepts, it needs to know the extent to which that improvement is comparable to the progress of equally able students at other institutions.

There are at least three arguments against using the value-added approach as a way to judge the contribution of the college or university to student learning that must be considered. One is that intellectual mastery at the end of college is overwhelmingly determined by students' preparation when they enter. However, as Stephen Klein has demonstrated in looking at the results of the CLA, some institutions clearly improve the higher-order intellectual skills of their students more than other institutions serving similar students do.

A second argument is that there are ceiling effects on student learning growth in elite institutions: there cannot be as much growth in institutions that admit students with high ability as measured, for example, by the SAT or ACT. If true, this would mean that the value-added approach
penalizes those institutions when compared to ones that admit students of lower ability. (Parenthetically, the reverse is the case now. Because colleges and universities neither use the value-added approach nor measure student learning outcomes in a comparative manner, only those that admit students with high SAT/ACT scores are considered to be of high quality.) But the ceiling is in fact low enough: students at the highest SAT/ACT levels only score fifty percent of what is possible on standardized tests that use the value-added approach. Moreover, elite institutions are free to compare the absolute level of student attainment they achieve, the scale they excel on, to that of their peers.

This, however, raises the third argument: that institutions, selective and non-selective alike, may want to assert their contribution to students' intellectual growth not by measuring value added but by setting performance standards. But institutions, even while doing value-added assessment, are free to set standards for minimum proficiency and low to high ranges in performance. Indeed, the producers of testing instruments could provide them with sample answers for well-below, below, average, above-average, and well-above-average performance.

**Why the Trend Toward Greater Transparency is Irreversible**

The assessment and accountability debates are not new, but there appears to be a new coalescence of actors pushing for more assessment of student learning, transparency with regard to university operations, and accountability: local government, state legislators and governors, state boards of higher education, Congress, the U.S. Department of Education, journalists, parents, and foundations. They are responding to rising college costs, the nation's need for human capital development, and the imperative that we do better in ensuring successful educational attainment by minority students, among other motivators.

So as the demand for greater transparency and accountability becomes an imperative for all institutions, one of the most important issues for us to address is whether and how information about higher education institutions should be reported publicly. We may be entering a period when we have comparative measurement instruments of sufficient quality to use for assessment and accountability purposes, but we are only in the beginning stages of serious discussion of their public policy implications.

For instance, as institutions, associations, systems of higher education, and states devise templates for reporting, they will publicize them. But raw scores on assessment measures may vary from one year to the next, low graduation rates may signal that an institution is doing a good job of recruiting at-risk students, and so on. The rules of engagement must be worked out through trial-and-error efforts led by national associations, systems of higher education, states, and individual institutions over the next several years.

We should apply to any proposed strategies for assessment and accountability the same logic by which Vannevar Bush developed the peer review research policy that respects the diversity of institutions and American higher education's institutional design, with its decentralized governance structure and respect for faculty autonomy. But we need to engage in the kind of layered comparison strategy described here to improve the quality of teaching and learning in higher education. If we did not have the ability to do the kind of comparative assessment described here, the status quo might be acceptable. But we now can do sophisticated, meaningful, appropriate comparisons—and, since we can, there is no legitimate argument against doing so. 

---

**Resources**