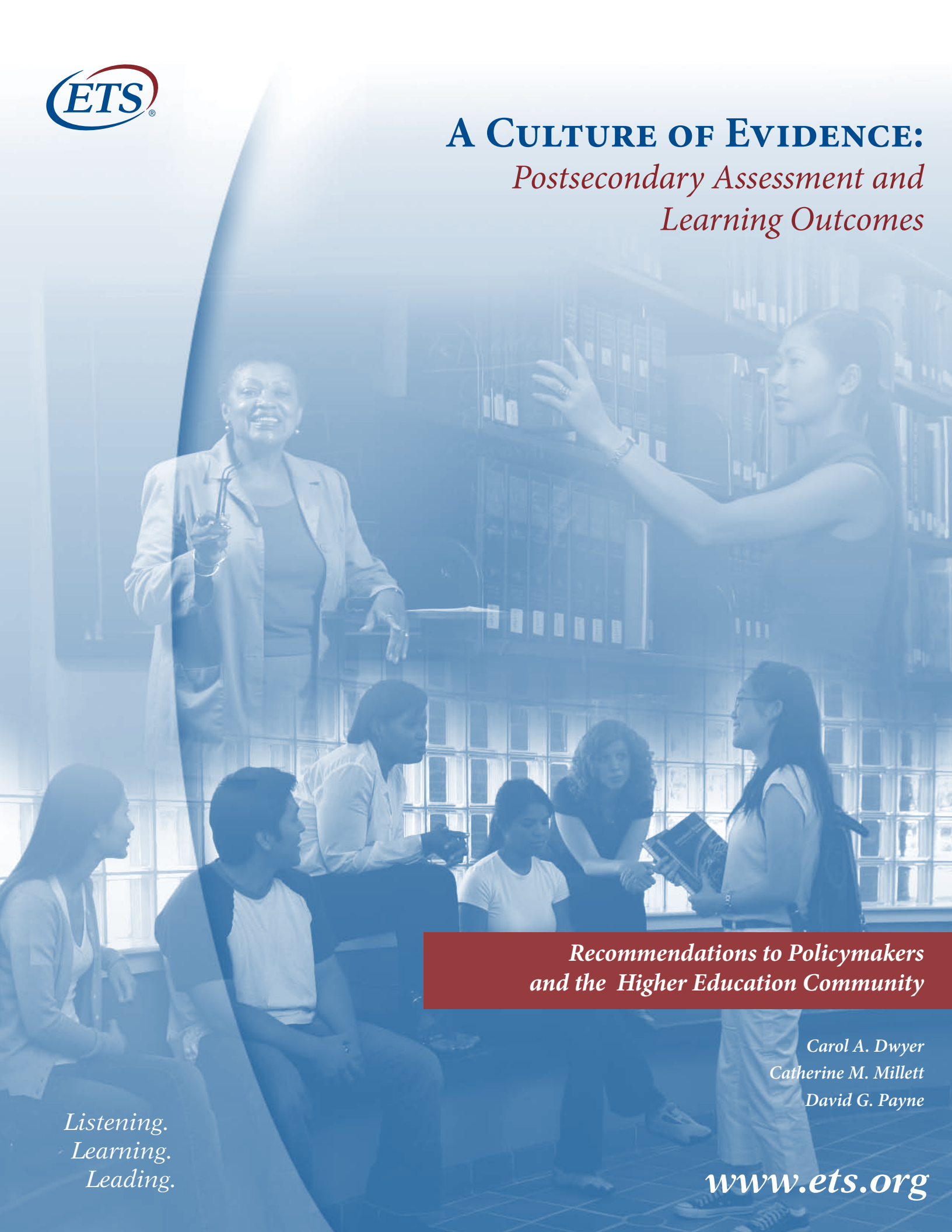
A Culture of Evidence:

Postsecondary Assessment and Learning Outcomes



Recommendations to Policymakers and the Higher Education Community

Carol A. Dwyer
Catherine M. Millett
David G. Payne

Listening.
Learning.
Leading.

www.ets.org

A Culture of Evidence:
Postsecondary Assessment and Learning Outcomes

Recommendations to Policymakers and the Higher Education Community

CAROL A. DWYER
CATHERINE M. MILLETT
DAVID G. PAYNE

ETS
PRINCETON, N.J.

June 2006

Dear Colleague:

Developing a comprehensive strategy for postsecondary education that will meet the needs of America's diverse population and help ensure our ability to compete in the global economy is vital to the growth of our nation.

The bar is being raised for the nation's higher education system. Americans realize that pushing students through the system is not enough; students must graduate equipped with the skills and knowledge needed to be productive members of the workforce.

Key to improving the performance of our colleges and universities is measuring their performance. Therefore, I am pleased to share with you this ETS issue paper titled *A Culture of Evidence: Postsecondary Assessment and Learning Outcomes*, which outlines accountability models and metrics for the higher education arena.

In this paper, we assert that to understand the value added to student inputs by the college experience, it is essential to address three measurements: student input measures, student output measures, and a measure of change between inputs and outputs. The paper also briefly reviews principles of fair and valid testing that pertain to the assessments being recommended.

Today's higher education institutions must not only prove their programs' performance; they must also take their programs to the next level if they are to be able to choose from the most promising applicants, attract prestigious faculty, and secure access to financial support from a competitive funding pool. Accordingly, colleges and universities should be held accountable to multiple stakeholders, ranging from students and parents, to faculty and administrators, to accreditation bodies and federal agencies.

As we move forward as a nation to improve postsecondary outcomes, I believe that the ideas set forth in this paper will help inform the national discussion on how we can improve our system of higher education.

Sincerely,

Mari Pearlman
Senior Vice President, Higher Education
ETS

Table of Contents

Executive Summary

Postsecondary education today is not driven by hard evidence of its effectiveness. Consequently, our current state of knowledge about the effectiveness of a college education is limited. The lack of a culture oriented toward evidence of specific student outcomes hampers informed decision-making by institutions, by students and their families, and by the future employers of college graduates.

What is needed is a systemic, data-driven, comprehensive approach to understanding the quality of two-year and four-year postsecondary education, with direct, valid and reliable measures of student learning. Most institutional information that we have access to today typically consists of either input characteristics (student grades and test scores, for example) or output characteristics (institutional counts of degrees granted or students employed, for example), with little attention to the intervening college-learning period.

We propose a comprehensive national system for determining the nature and extent of college learning, focusing on four dimensions of student learning:

- Workplace readiness and general skills
- Domain-specific knowledge and skills
- Soft skills, such as teamwork, communication and creativity
- Student engagement with learning

To understand the value that a college experience adds to student inputs, three measurements must be addressed: Student input measures (What were student competencies before college?), student output measures (What were student competencies after college?), and a measure of change between inputs and outputs.

This paper also briefly reviews principles of fair and valid testing that pertain to the assessments being recommended. The design for these measurements must include attention to the following points:

- Regular (preferably annual) data collection with common instruments
- Sampling of students within an institution, rather than testing all students, with an option for institutions that wish to test more (the unit of analysis is thus the institution)
- Using instruments that can be used in pre- and post-test mode and that have sufficient forms available for repeated use over time
- Using a variety of assessment formats, not limited to multiple-choice
- Identifying appropriate comparisons or "peer groups" against which to measure institutional progress

The paper concludes that there are currently no models or instruments that completely meet the needs of a comprehensive, high-quality postsecondary accountability system as outlined here.

We recommend that the six regional postsecondary accrediting agencies be charged with integrating a national system of assessing student learning into their ongoing reviews of institutions.

To consider moving in this direction, policymakers and the higher education community may wish to:

- Focus on early implementation of measures of workplace readiness and general skills.
- Convene an expert panel to review an Assessment Framework Template included in this paper.

- Charge the panel with reviewing the dimensions of learning to reach consensus on a framework; review the completeness of the list of extant assessments; and review each assessment to determine its match to desired skills and its applicability to both two-year and four-year institutions.

A detailed list of issues for consideration by such an expert panel is included.

> *To send your child off to a $40,000-a-year school, you just get "the feeling." Asked whether Mary's college is getting the job done, [Mary's mother] says: "The truth of the matter is, I think it's good but I have no way of knowing that — that's my point. She seems happy. For this kind of money she ought to be." (Toppo, 2006).*

This mother's appraisal of our current state of knowledge about the effectiveness of a college education in general or at a particular institution is most likely shared by students, other parents, government officials, business leaders, and future employers of college graduates. The public's knowledge about what happens once students start a college education is limited. We often make assumptions about the quality of an education based on the institution's reputation, and one occasionally hears statistics about college graduation rates. But what hard evidence is consistently available about the outcomes of a college education? The simple answer is there is no commonly used metric to determine the effectiveness — *defined in terms of student learning* — of higher education in the United States.

As we outline what a new era in higher education accountability might look like, we will strive to keep in mind two points: the need for clarity and simplicity in the system; and the need for a common language that can be used consistently within the higher education community as well as with stakeholders outside this community.

What is the purpose of a college education? Is it a first step toward advanced study? Is it for getting a better job? Is it preparation for being a better citizen and contributing member of society? Has there been a disconnect between education and work? Students are admitted to colleges and universities, complete courses, graduate, and then enter the world of work. But are they prepared for what employers expect them to know and be able to do? Whose responsibility is it to provide answers to these questions?

The three major players in accountability are the legislative and political arenas, the academy, and the general citizenry (LeMon, 2004, p. 39). They all need reliable and valued information in a useable form. We must ask: What have students learned, and are they ready to use it? (Malandra, 2005).

When the National Center for Public Policy and Higher Education awarded all 50 states an "incomplete" in the student learning category in its 2000 inaugural issue of *Measuring Up*, the higher education community, policymakers and the public got their first inkling of the paucity of information about student learning in college. Miller and Ewell (2005) took a first step in framing how individual states might begin the process of measuring student learning outcomes by considering several data-oriented themes: (a) the literacy levels of the state population (weighted 25% in their overall evaluation); (b) graduates' readiness for advanced practice (weighted 25%); and (c) the performance of the college-educated population (weighted 50%). To get the process started, Miller and Ewell's college-level learning model employed currently available assessments. For example, literacy levels were assessed using the 1992 National Adult Literacy Surveys, now known as the National Assessment of Adult Literacy, or NAAL, which poses real-world tasks or problems for respondents to perform or solve (2006). The graduates' readiness for the advanced practice section used extant data on licensure examinations, competitive admissions exams, and teacher preparation exams. The most heavily weighted component, performance of the college educated, analyzed student performance on the ACT Workkeys assessments for two-year institutions and the Collegiate Learning Assessment (CLA) for four-year institutions.

Two of the assessments of college-level learning in *Measuring Up* warrant additional comment. One NAAL finding in particular caught the public's attention: "only 31 percent of college graduates can read a complex book and extrapolate from it" (Romano, 2005). The CLA has also continued to be in the public eye. Interest in the CLA may be due to several of its appealing qualities: institutions rather than students are the unit of analysis, pre- and post-test measures can be conducted, and students construct their own responses rather than answer multiple-choice questions. According to *CLA in Context 2004-2005*, approximately 134 colleges and universities have used the CLA since 2002 (Council for Aid to Education, 2005).

At approximately the same time that *Measuring Up* was building momentum, the National Survey of Student Engagement (NSSE) was in development. Begun in 1998, NSSE has collected information about student participation in programs and activities that promote learning and personal development. The survey provides information on how college students spend their time and their participation in activities that have been empirically demonstrated to be associated with desired outcomes of college (NSSE, 2005a). The information thus represents what constitutes good practices in education. Although the data are collected from individual students, it is the institutions rather than the students that are the units of analysis. Over 970 institutions have participated in NSSE and new surveys have been developed for other important sectors such as law schools (LSSSE), community colleges (CCSSE), and high schools (HHSSE).

The project described by Miller and Ewell (2005) and the assessments of student engagement represent two of the recent efforts to answer questions regarding institutional effectiveness in U.S. higher education. To appreciate the contributions of these efforts, and to provide a framework for the present proposal, it is important to review briefly some of the major characteristics of U.S. higher education at the start of the 21st century.

Access for all is the hallmark of the U.S. postsecondary education system. As a nation, we are justifiably proud of the fact that a college education is possible for all citizens, ranging from the traditional high school graduate to the senior citizen who wishes to fulfill a lifelong dream of earning a college degree. Another important facet of U.S. higher education is the relatively large degree of autonomy given to institutions of higher education (IHEs). Similarly, faculty are often given tremendous autonomy in setting the curriculum, establishing degree requirements, and other important academic matters. These aspects of U.S. higher education represent important contextual features of the organizations that must be kept in mind as we consider new accountability measures, especially for student learning outcomes.

In addition to broad access and institutional autonomy, other aspects of U.S. higher education provide a lens through which to view the state of affairs in higher education; that is, important dimensions along which institutions can be described. Although numerous discrete dimensions may be used (e.g., public vs. private, for-profit vs. nonprofit, two-year vs. four-year, selective vs. nonselective), more nuanced dimensions provide a richer set of descriptors.[1] The image of a series of continua is most appropriate for thinking about the U.S. system. Some examples illustrating these continua and their underlying complexity can be usefully considered from the institutional, student and learning environment perspectives. These dimensions are important for present purposes because they relate directly to approaches that can be used to assess student learning for the purposes of monitoring and improving institutional effectiveness in the teaching and learning domains.

Institutions

- Postsecondary institutions award academic credentials ranging from certificates to doctoral degrees.
- The instructional level of institutions ranges from less than one-year to four-year.
- The degree of selectivity differs greatly among institutions.
- There are several sectors within the postsecondary level (e.g., public vs. private, and nonprofit vs. for-profit).
- Postsecondary institutions differ in their histories and institutional missions, (e.g., religiously oriented institutions, Historically Black Colleges and Universities, Hispanic Serving Institutions, Tribal Colleges).
- Institutions range from being highly centralized to highly decentralized.
- In 2002, the 4,071 U.S. postsecondary institutions ranged in size from those enrolling fewer than 200 students to those that enrolled 40,000 or more (NCES, 2002a).
- In 2001, nearly 16 million students were enrolled in U.S. degree-granting institutions. Public institutions enrolled 77% of all students; private nonprofit institutions enrolled 20% of students; and private for-profits enrolled 3% of students (NCES, 2002b).

Students

- Students range from traditional age (recent high school graduates) to older adults. In 2001, 37% of students enrolled in four-year and two-year institutions were 25 or older (NCES, 2002b).
- The number of institutions that a student attends can range from one to four or more. A majority (59%) of all of the 1999-2000 college graduates (first-time bachelor's degree recipients) attended more than one institution (Peter & Forrest, 2005).

- Looking only at traditional-age students, between 54% and 58% of those who started in a four-year college earned a bachelor's degree from the same school within six years of entry. For those who earned a degree from a different four-year college than the one in which they began, the six-year completion rate is between 62% and 67% (Adelman, 2006).

The Learning Environment

- Universities employ a range of selection criteria for admitting first-year students.[2] For example, 83% of public four-year and 72% of private non-profit four-year institutions review admissions test scores, compared with only 4% of two-year public institutions (*Chronicle of Higher Education*, 2005).
- More than a quarter of entering first-year students in fall 2000 enrolled in at least one remedial reading, writing or mathematics course (Parsad & Lewis, 2003).
- The learning environment takes many forms today, ranging from a faculty member lecturing behind a podium on Monday mornings to faculty teaching an online course that students can access at any time to suit their own schedules.
- Course offerings range from complete centralized standardization of content to near-total faculty control of the content.
- Institutions differ in their perspectives on what every student should know. At one extreme is Brown University, which has no core requirements; a general-education requirement is in the middle; and the great-books-style curriculum of Columbia University and the University of Chicago is at the other extreme (McGrath, 2006).
- The most popular disciplines for associate's and bachelor's degrees combined are business (20%); liberal arts, sciences, general studies and humanities (13%); health professions and related clinical sciences (8%); and social sciences and history (8%) (NCES, 2002c).

The dimensions of institutional characteristics, the nature of the students who apply to and enroll in colleges and universities, and the learning environments created in these institutions are all critical aspects of the U.S. higher education system. As such, they must be taken into account as we contemplate the creation of a system of accountability for student learning. In the next section we will introduce a conceptual model that organizes these dimensions as they relate to the primary function of colleges and universities — teaching students.

The I-E-O Model

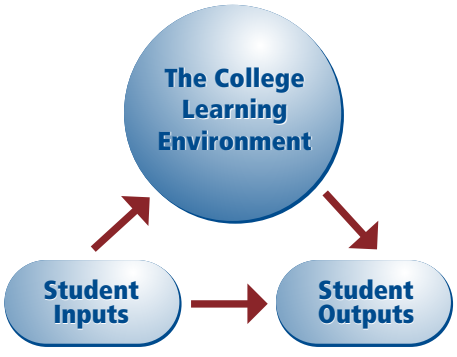
One approach to thinking about higher education is to embrace an econometric model examining inputs and outputs. Such an approach has a number of merits, including the need for careful articulation of the inputs into the system, the resources invested in the system, and the outcomes produced by the system. Such approaches can yield important insights into areas such as efficiency. For example, for every $100,000 invested in public higher education in a given state, how many graduates are produced? For every 100 students who enter the system, how many are retained in the second year, and how many graduate within four years?

A shortcoming of a strict econometric approach to studying higher education is that it can run the risk of ignoring one of the most important, yet difficult to measure, functions of colleges and universities: facilitating students' learning.[3] Because the focus of this paper is on student learning, it is important to deal with students as they pass through various stages in the education system, for example, completing the first year or earning an associate's degree. The primary interest is in the means that can be used both to characterize and to understand the learning that takes place in colleges and universities. Having knowledge of the passage of students through the system is a necessary but not sufficient aspect of accountability in higher education: If accountability were restricted to measures of, for example, retention rate from the first to the second year, or graduation rate, then it would be possible to improve on these measures in ways that would not necessarily increase or improve student learning. For instance, lowering standards for grading could increase retention and graduation rates, but it might well hinder student learning.

Given this educational context, a slight modification to Astin's simple yet elegant input-environment-outcome model serves as a framework for thinking about the overall college experience from both an institutional and individual student perspective (see Figure 1). The proposed student input-college learning environment-student outcome model illustrates that student inputs — for example, the intensity of high school academic preparation — can have a direct relationship to student outputs, such as degree completion (Adelman, 2006). It also illustrates that the college-learning environment can have a direct effect on student outcomes.

Figure 1

The I-E-O Model



Source: Astin (1993).

The Institutional Perspective

If one perspective on the collegiate experience takes institutions as the unit of analysis, then it is important to consider what information is currently and consistently available about U.S. institutions. Most of the institutional information to which the direct consumers of a college education (students) and the indirect consumers of higher education (employers, government and communities) have access is based on two principal metrics — student performance and institutional counts. These metrics typically consist of either input or output characteristics, with little accounting for what happens in the intervening period — the college-learning period.

First consider the inputs. There are two basic types of input measures: measures of simple quantity, such as the number of applicants or the number of students admitted; and measures of quality or academic preparation. Some common postsecondary institutional input measures include average performance on the SAT/ACT, average high school grade point average, the number of National Merit Scholars, the number of students who have advanced standing[4], and institutional admission yields (percentage accepted who enroll). From the perspective of the consumer, institutional characteristics to which students may have access when they are considering applying to college pertain to the institution as a whole — the size of the student body, its academic reputation in rankings such as *U.S. News and World Report* or *Barron's Profiles of American Colleges*, the faculty's academic credentials, the number of faculty who have won prestigious prizes in their fields, the size of the library collection, and the size of the institution's endowment.

At the other end of the education experience, there are two typical classes of output measures for educational institutions. As with input measures, these can be broken down into quantity and quality measures. For example, institutions report the number of degrees granted. The characteristics of degree recipients tend to be reported as average student performance on graduate and professional school admissions tests, such as the GRE, GMAT, LSAT and MCAT; performance on licensure exams, such as the NCLEX and *Praxis Series* assessments; and the percentage of students with jobs after graduation or plans to enter graduate or professional school. One of the potential limitations of these data is that they are not representative of the entire student population. For instance, only students who are interested in attending medical school typically take the MCAT. Another limitation is that standardized college admission measures are available for less than half of today's two- and four-year college students. Underrepresented minority students, because they attend community colleges in large numbers, are disproportionately among those not having taken these tests.

The Student Perspective

A second perspective on the collegiate experience is that of the students. The reality is that students enter the U.S. postsecondary system with varying stockpiles of academic accomplishments and skills. Consider two students who enter college with the same goal of one day earning a bachelor's degree in economics. At face value, the only difference between these students is that Student One enters college with a 36 ACT/1600 SAT score, a 4.0 GPA, and a score of 5 on six AP exams; while Student Two enters college with a GED and the need to take remedial mathematics, writing and reading. Student One earned a bachelor's degree in economics in four years and entered the workforce; Student Two earned a bachelor's degree in economics in six years. These two students share two elements in common — they both have bachelor's degrees and their majors were economics. What we do not know about our students are the following:

- Do they have the same knowledge of economics?
- Are they both ready to enter the workforce?
- What type of engagement did they have with their collegiate experiences?
- What type of "soft skills," such as ability to work in teams, do they have?

What does it mean if the students differ with respect to these questions? What is the consensus on whether bachelor's degree recipients should share some common achievements?

The example above could be considered in many different contexts. For example, we could take a group of students with a 36 ACT/1600 SAT score, a 4.0 GPA, and a score of 5 on six AP exams, and randomly assign them to different institutions where they can major in economics. We could then ask the four questions above to determine if their outcomes are similar. The answers to these questions might help the mother quoted in the introduction have better information on where her child ought to apply to college.

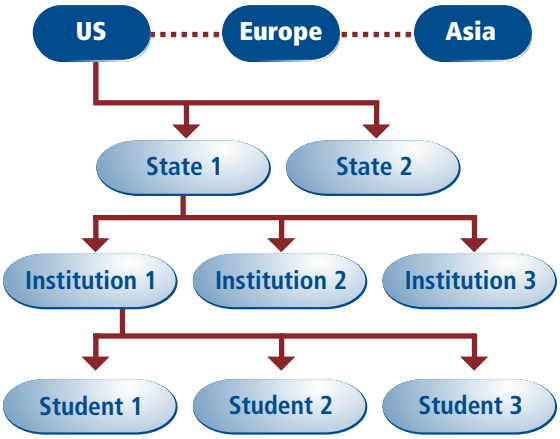
Within the U.S. education system, there are many different forms of postsecondary peer groups (see Figure 2). These peer groups are salient to different institutions and stakeholders in different ways: some exist primarily for historical reasons and some for practical reasons such as competition for market share. An important element of institutional peer groups for current purposes is establishing comparability among institutions that are to be compared. A college may wish to benchmark its performance relative to a set of self-defined peer institutions, or a set of "stretch" comparisons with institutions that represent the next level the institution aspires to reach. Peer group comparisons are also useful in a global education marketplace, allowing, say, the U.S. postsecondary system to be compared with systems in other countries. European Union members are currently developing a set of descriptors of the knowledge that would represent mastery of given academic domains that will compare the different education systems within the EU. International comparisons may gain in importance as the global race for talent intensifies.

In the United States, there are at least three other types of peer groups that might profitably be studied. First, states are a natural peer group in the U.S. policy arena. Second, institutions often have peer institutions against which they benchmark their accomplishments. Some well known examples are the Ivy League, the Big Ten, and the PAC Ten, but almost all institutions have some form of peer group that is of use to them in their institutional decision making. And at the student level, research often organizes students into peer groups on the basis of prior academic achievement, gender, race/ethnicity, and income.

After identifying peer groups, the next issue to address is how to use the performance of these peer groups. In the case of student learning outcomes, we are interested in assessments that provide an index of student learning. To use these assessments for purposes of accountability and improving U.S. higher education, it is essential that these assessments enable us to make appropriate comparisons and draw conclusions based on these comparisons. The following sections summarize some assessment characteristics relevant to assessment for postsecondary accountability and improvement.

Figure 2

The U.S. education system in an international and national context



Carefully designed assessments address fundamental concepts and attitudes that are key to improving higher education: operationalizing accountability and personal and organizational responsibility; increasing the rationality, fairness and consistency of comparisons; enhancing accuracy, replicability and generalization; providing data for description and prediction; and fostering understanding of personal and institutional exceptionalities, both positive and negative. In the context of increasing the value added to student learning by higher education, appropriate measurement can help highlight more specifically where problems are and how severe they are. Measurement can also help identify success stories that illustrate what actually works and how to apply that success to other settings.

The gold standard for judging the quality of assessments of the type discussed in this paper is the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The *Standards* contain comprehensive, widely accepted, and highly detailed guidance for fair and valid educational testing; a full treatment of them would be well beyond the scope of this paper. It is important to note, however, that the entire set of standards is based on the premise that validity refers to the degree to which evidence supports the interpretation of test scores, and that this is the most fundamental consideration in assuring the quality of any assessment. In their simplest form, there are four essential validity questions: 1) How much of what you want to measure is actually being measured? 2) How much of what you did not intend to measure is actually being measured? 3) What are the intended and unintended consequences of the assessment? and 4) What evidence do you have to support your answers to the previous three questions?

This is the modern view of assessment validity (see, e.g., Messick, 1989), and it underscores the importance of examining tests within their total context. This means that the validity of a test can no longer be conceived of as simply a correlation coefficient, but rather as a judgment, over time, about the inferences that can be drawn from test data, including the intended or unintended consequences of using the test. For this reason, good test design and use explicitly include many considerations beyond the test itself. We need to be able to infer from a test score that it accurately reflects student learning. For this inference to be valid, we must assume that individual test takers are motivated to put forth sufficient effort to demonstrate their actual knowledge and skills. Good assessment design thus requires eliminating this threat to validity through appropriate attention to incentives to students to participate meaningfully. This is an issue that will be of great significance in any postsecondary accountability assessment.

This comprehensive view of good assessment provides a means to ensure fairness as well. Valid assessment requires clarity and completeness in specifying what an assessment is and is not supposed to measure, and requires evidence of this for *all* test takers. This means that a test that is not fair to some of the test takers cannot be valid overall. Tests that show real, relevant differences are fair; tests that show differences unrelated to the purpose of the test are not.

It is most useful for present purposes to consider assessment as a comprehensive, iterative cycle of measuring progress at multiple points in time, using the resulting data to understand problems and to design, and ultimately implement, effective curricular improvements in higher education.

In addition, higher education is a complex set of levels and participants formed into peer groups within which useful comparisons can be made. We envision the need to define such groups as consisting of individuals, institutions, states and nations. Useful and valid assessments will have to take these groups and their interactions into account. Assessment data will necessarily be multidimensional, and will reflect individuals' and institutions' perspectives and

needs. This complex picture should focus on rationally defined sets of comparisons on specific dimensions, which will necessarily preclude simple, uni-dimensional ranking schemes.

The American Educational Research Association (AERA), one of the sponsoring organizations of the *Standards*, has issued a position statement on high-stakes testing in the realm of elementary and secondary school settings that has relevance for higher education as well (AERA, 2000). Specifically, it makes the following important points:

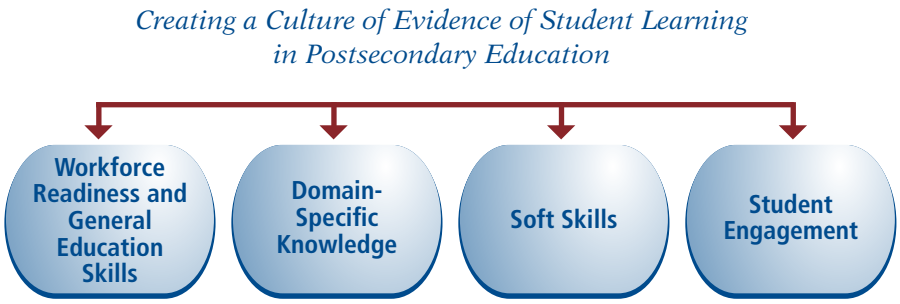
- High-stakes decisions should not be based on a single test.
- Testing to reform or improve current practice should not be done in the absence of provision of resources and opportunities to improve learning.
- Tests that are valid for one use or setting may not be valid for another.
- Assessments and curricula should be carefully aligned.
- Establishing achievement levels, categories or passing scores is an important assessment activity that needs to be validated in its own right.

In the rest of this paper, we make the assumption that these features of fair, useful and valid assessment will be part of the postsecondary assessment activities we propose.

There is a growing consensus among educators and business leaders that student learning in higher education is multifaceted and that it therefore needs to be assessed using different tools for the major learning dimensions. Individual institutions of higher education need to assess the extent to which they are succeeding in meeting the highly specific learning objectives that align with their particular missions.[5] At the state and national levels, however, there are common themes in student learning at the postsecondary level. In this section we summarize three major dimensions of student learning that could be assessed across public two- and four-year postsecondary institutions, and a fourth dimension, student engagement, that is important to students' success and should be carefully monitored, but is not in itself a student learning outcome (see Figure 3).

Figure 3

Domains of Student Learning



1. Workplace Readiness and General Education Skills

To succeed in the workforce or to proceed to higher levels of academic or professional performance, learners must acquire a set of basic minimum skills and abilities. Academic and business leaders have identified a set of abilities for which there is wide agreement about importance. These include: (a) verbal reasoning; (b) quantitative reasoning, including basic mathematic concepts such as arithmetic, statistics and algebra; (c) critical thinking and problem solving; and (d) communication skills, including writing. These basic academic skills are taught in a variety of courses across the undergraduate curriculum. These skills are also taught in the full range of institutions of higher education, regardless of degree level (e.g., community college, four-year institutions), source of funding (public vs. private), or business objective (for-profit and nonprofit). These skill sets may be given somewhat different labels in different contexts, but at their core they reflect the skills and habits of mind that are necessary to succeed in both academic and workplace settings. As such, they merit close attention in any system of accountability for student learning.

2. Content Knowledge/Discipline-Specific Knowledge and Skills

To become a member of most professions, there is a set of knowledge and skills that one must acquire in order to be considered competent within that domain. Again, this fact applies across different types of postsecondary institutions and a broad range of degree levels (e.g., certificates, and associate's, bachelor's, master's and doctoral degrees). Many disciplines (e.g., health professions, law and business) also require professional-certification examinations that define the qualifications needed to enter the profession.

There is also a large number of academic disciplines, especially in the arts and sciences, in which there are no certification standards. In these areas, in lieu of such standards, the awarding

of the degree or certificate is taken as evidence of mastery of the core set of competencies. The assumption is thus that a person with, say, a bachelor's degree in chemistry is expected to be familiar with and able to use a common core of knowledge of organic chemistry, inorganic chemistry, and other subcategories.

As education plays a more central role in determining nations' economic well-being, measures of learning success beyond basic acquired academic abilities will continue to loom large. The importance of discipline-specific knowledge and skills has been acknowledged by leaders of the education reform movement in Europe (For a summary, see *http://www.ntb.ch/SEFI/bolognadec.html*). With the increasingly mobile student population across Europe and the movement to standardize systems of credits and degrees, European educators have begun work on a set of descriptions of the minimum competencies expected in the major academic disciplines (Joint Quality Initiative, 2004). These so-called Dublin descriptors will be important to members of the European education community as they create national education systems and policies that allow students to move across national boundaries as they pursue their education and vocational goals.

As state and federal leaders continue to ask increasingly urgent questions regarding the return on investment in higher education, it is critical that they consider more than just the broad classes of learning typically identified with General Education requirements. By asking the extent to which students are becoming proficient and knowledgeable in their chosen fields, we can further the dialogue on education quality and improvement. As with the other dimensions of student learning, it is essential to have a system of assessment that allows comparisons across various benchmark groups, including national, state, regional and peer groups.

3. "Soft Skills" (Noncognitive Skills)

In today's knowledge economy, it is not sufficient for a worker to possess adequate basic cognitive skills and discipline-specific competencies. The nature of work also requires that the person be able to work in teams, be a creative problem solver, and communicate with a diverse set of colleagues and clients. Employers, colleges and universities have become more cognizant of the role that such so-called "soft" or noncognitive skills play in successful performance in both academic and nonacademic arenas. The measurement of skills and traits such as creativity, teamwork and persistence has become a major focus in applied areas such as human resources and industrial-organizational psychology. The importance of noncognitive skills is well established within academic settings, but there are fewer widely adapted approaches to measuring these skills in education settings than there are in the industrial, governmental and nongovernmental domains.

4. Student Engagement

In addition to assessing the three dimensions of student learning, it is also appropriate to ask questions regarding the extent to which best education practices are reflected in the education system, and the extent to which students are actively engaged in their own learning. As mentioned earlier, a great deal of information on student engagement has already been amassed by NSSE; additional details of this effort are given below.

In recent years, there has been a growing scientific and social recognition that students play an active role in their own learning, and that any attempt to characterize the learning that takes place in higher education must consider the individual student's role in this process. At four-year institutions, NSSE (2005b) was created in 2000 with support from the Pew Trusts. In the spring of 2005, 529 colleges and universities participated in the survey, which represents approximately one-quarter of the four-year colleges and universities in the United States. In addition to

information about what IHEs are offering their students (from the students' point of view), NSSE also collects information that is related to students' own efforts to learn. For example, there are questions about how much homework a student does in a typical week, how much time was spent socializing, and how much reading the student did beyond that which was assigned. Student engagement has also been assessed at community colleges through the Community College Survey of Student Engagement (NSSE, 2005a). In 2005, 257 colleges participated in CCSSE.

It is important to understand that student engagement is not, in itself, an index of student learning. Rather, student engagement is an index of the nature and extent of the student's active participation in the learning process, and NSSE and CCSSE are intended to measure what students do in school. These surveys do not provide independent measures of the learning that is assumed to take place as a result of these activities. Student engagement is, however, considered by many to be both a valuable aspect of postsecondary education for the individual and the institution, and an indicator of motivation and habits that carry over into other current and future settings.

To summarize, we have identified three important areas of student learning that can be assessed at all institutions of higher education, and one domain that concerns the academically related activities that students engage in during their undergraduate careers. At present, there is no data set that provides comprehensive data for the three measures of learning or student engagement. Although there are a number of measures available, including both standardized assessments and locally developed measures, there is no overall data set that would allow legislators, policymakers, students, parents and employers to obtain detailed information about student learning and student engagement. We believe that continuing to collect data in these areas would provide stakeholders with valuable information.

We now move to a discussion of the nature of the data that would have to be collected in each of these domains, including the sampling procedures, the relations among the various measures, and how these data can be used for benchmarking purposes.

Broadly speaking, there are three points in time at which we can assess student learning, and one derivative measure that can use data from the first three types of assessment. First, we can evaluate the competencies of students who apply to and enroll in our colleges and universities. Second, we can assess their performance as they progress through their degree programs. Third, we can evaluate what they have learned by the time they graduate, or when they are about to enter the workforce or graduate school. Within higher education, these three types of measures correspond, respectively, to admissions or placement measures; formative or summative assessments completed within the curriculum but before completion of the degree; and outcome measures or admission measures for graduate and professional school. Finally, we can gauge the "value-added" aspect of their education by comparing an initial measure of competency with a measure taken after the student has completed some of their entire intended degree program.

Each of these four classes of assessment has utility for one or more purposes. They also each exhibit some limitations as an index of student learning when used in isolation. For example, if an institution decided to measure student learning using only an outcome measure (e.g., performance on a standardized test of general education competencies or scores on a graduate admissions examination), then this would tell the institution something about how well prepared their students were for the workplace or graduate school. But we could not infer that the levels of performance on these measures reflected only the impact of the institution: a highly selective undergraduate institution would be expected to have students who scored high on the GRE even if the institution had not contributed a great deal to the students' learning. Conversely, consider an institution that takes in a diverse set of students who are not strong academically at the start of college (e.g., those scoring at the 10th percentile on the ACT). If the institution does an outstanding job with these students, but at the end of the students' undergraduate careers the students score at the 50th percentile on the GRE, does this mean that the institution has failed to produce graduates who are well prepared? Or does it mean that the institution has done an excellent job, as indicated by the fact that the students moved from the 10th to the 50th percentile? Alternatively, do the GRE data tell us only about that small portion of the students who have decided to continue on for graduate education? Have we learned anything at all about the accomplishments of the student body as a whole?

To gain an understanding of the range of contributions that institutions of higher education make to student learning, it is essential to consider *at the very minimum* three measurement points: an input measure (What were the student's competencies when they started the program?); an outcome measure (What did the students know and what were they able to do when they graduated?); and a measure of the change between these inputs and outputs. Depending upon the question of interest, one or more of these three measures may be of primary interest, but it is essential to note that without measurement at all three points, it is impossible to fully comprehend student learning and the contributions of an academic institution to this learning.

Another key aspect of the data that are needed to evaluate student learning is the comparability of the measures collected. In simplest terms, it is essential that the various assessments allow one to compare the measures in ways that are not flawed. If the input measure were a standardized test of reasoning skills and the output measure were a locally developed survey of student engagement, it would be impossible to make any direct comparisons that reflect student learning. This is not to say that the data from these individual measures are without value. Information from standardized admissions tests can inform as to the selectivity of the institution, the quality of enrolled students, and other important factors. Measures of student engagement can tell academic leaders about how widespread various forms of teaching are, how much students from various groups interact with faculty, how much time students say

they spend studying, and so forth. But we could not use these data to make any inferences about how much students had learned as a result of their classes or their interactions with faculty and classmates.

Returning to our review of the minimum data necessary to measure student learning, we need to point out that even with data from these three measures in hand (input, output and change), we are still lacking a critical piece of information. A simplified thought experiment will demonstrate the need for another type of measure. Imagine that, on average, students at University X show an increase of 10% in a particular competency domain, for example, writing ability over the course of their undergraduate careers. How should the increase be interpreted? Is it attributable to the courses that students take at University X? To know this would require identifying a control group of students who were comparable to the students at University X in all important respects but who did not take any courses between the time when the input and output measures were taken. For example, we might run a controlled experiment in which we randomly assigned students to attend either University X or not to attend a university at all. This would allow a type of causal inference to be drawn regarding the effects or impact of an education at University X on student learning.[6]

This thought experiment illustrates another type of data that is needed to interpret the contribution of education programs to student learning validly, namely, a means by which we can compare relative changes in student learning across different comparison groups. To continue the previous example, if University X had identified a set of 10 peer institutions that could be considered similar to University X (e.g., similar incoming student credentials, size of institution), then having comparative data from these 10 institutions would allow the leaders of University X to determine if they were doing better than, worse than, or about the same as the set of peer institutions.

The final aspect of student learning data that is necessary to consider is the historical nature of these data. It is important for any quality improvement initiative to have data collected at regular intervals in order to track progress. Depending upon the complexity and magnitude of the data to be collected, along with the specific needs of the institution and any state or federal initiative that is driving the specific data collection effort, it may be possible to consider various sampling plans. Ideally, data would be collected on an annual basis, and then analyzed on a schedule that met the needs of the organization(s) using the data.

Summary

We have identified four critical dimensions to be considered in designing assessment of student learning and the learning process. There are three aspects of student learning domains — general education, discipline specific, and soft skills, along with self-reported survey information about student engagement; a minimum of three types of measurement, including input, output and a derivative change measure that indexes the impact of learning attributable to the educational institution; and the necessity of regular, preferably annual, data collection efforts. In addition to these aspects of measuring student learning, there are issues to be considered regarding the appropriate units of analysis (e.g., state level data, sector data, national data) and comparisons to be made.

If the United States is seriously interested in measuring student learning in postsecondary settings, then there must be a national initiative to create a system for collecting data. Ideally, that system would be capable of examining performance in the three learning categories identified earlier — general education; discipline-specific skills and knowledge; and soft skills — as well as measures of student engagement. We have already identified the main features that such an assessment system must possess; in this section, we set forth a specific set of recommendations.

Our recommendations build upon the excellent work done by the National Center for Public Policy and Higher Education and reported by Miller and Ewell (2005). As noted in our introduction, this effort was directed toward providing answers to two broad questions:

1. What is the "educational capital," or the knowledge and skills of the population, that states have available to them for developing or sustaining a competitive economy and vital civic life?

2. How do all the colleges and universities in the state — that is, public, private, nonprofit and for-profit — contribute to the development of the state's educational capital?

The National Forum convened by the National Center for Public Policy and Higher Education addressed these two questions within the context of what can be done at the state level to examine state performance in higher education. As noted above, the Forum articulated three themes that can be examined when assessing how well states are doing in higher education: literacy level of the state's population; graduates who are ready for advanced practice; and the performance of the college educated population. These themes also provide a context for our current recommendations.

Our recommendations are concerned primarily with the third theme, which in the case of Miller and Ewell's work focused specifically on the question, "How effectively can students who are about to graduate from two- and four-year colleges and universities communicate and solve problems?" (Miller & Ewell, 2005, p. 9). We have gone beyond this question by recommending the design of an assessment system that would also answer the related question of how well institutions are doing, relative to their peer groups, in contributing to the learning of the students they enroll.

We will also go beyond the National Forum project by recommending a set of approaches that can be used to create a truly comprehensive, as opposed to a piecemeal, system of accountability for student learning. Our recommendations are developed within the context of the three primary dimensions of student learning that we have described in this paper (general, discipline-specific, and soft skills), along with a consideration of student engagement. Finally, after we describe our recommendations in some detail, we will comment on the remaining two themes identified by the National Forum.

Recommended Plan: A National Initiative to Create a System for Assessing Student Learning Outcomes in Higher Education

At its core, our recommendation is that the United States approach the development of a comprehensive system of accountability for student learning by using a phased strategy that, over time, will address each of the three major learning dimensions described above. In addition, and of critical importance, the recommendations offer the possibility for individual institutions to further expand the assessment of student learning in ways that are consistent with the specific mission of the institution and with the resources that are available to it.

Workforce Readiness and General Education Skills

We recommend that the United States invest now in the creation of a system that will assess the knowledge and skills that correspond to the dimension of workforce readiness. Because our proposed focus is on the value added by institutions, this system should employ a sophisticated sampling procedure analogous to that used with the National Assessment of Educational Progress. Although a sampling procedure would be central to the requirements of the national assessment of workforce readiness and general skills, this would not preclude individual institutions or states from considering a system that would employ a more comprehensive assessment that would include larger numbers, or even all, of their students in order to meet other, more institutionally tailored decision-making needs.

There are several reasons for our recommendation to begin this comprehensive system with a strong focus on workforce readiness: the economic importance of the skills; the existing consensus on them; and the available measures in multiple versions.

First, from an economic perspective this dimension has direct links to the economic competitiveness of the United States in a global economy. Workforce readiness skills are in many ways the foundations of skill development upon which higher-level skills are acquired, either at the undergraduate level with the attainment of a bachelor's degree, in advanced work in graduate school, in on-the-job training, or in lifelong learning in a range of contexts.

Second, there is a consensus on many of the main dimensions of skills needed to succeed in the workplace and in higher education. There are, of course, differences of opinion as to the exact details of the nature of these broad skill areas, and how best to measure them or teach them. But the fact remains that the consensus on these skills gives us a strong starting point.

Third, there now exist well-developed standardized assessment procedures that can be used to measure these skills. From the perspective of those in higher education, there is another feature of the currently available assessment tools that will help ensure their acceptance within the academy: many of these skills can be measured using constructed-response measures such as writing, as well as multiple-choice measures. It is important to note that the use of constructed-response measures is one of several features of the CLA that has helped it gain acceptance among academics.

Fourth, it is possible to create multiple, comparable versions of measures of workforce readiness so that we can measure skills among students entering institutions as well as upon graduation or after completion of a portion of the undergraduate curriculum. This allows us to measure the critical value-added aspect of student learning that is necessary when asking questions regarding how much IHEs are contributing to student learning.

Finally, despite differences in education systems around the world, workforce readiness is likely to play an important role in how employers, policymakers and educators evaluate how well education is "working" in countries around the world. As recent events have shown, multinational corporations are able to make decisions on where to invest resources based on, among other things, the availability of a well-educated workforce. An assessment system that provides clear information about workforce readiness and the contributions of colleges to the development of that readiness will be a major asset for future decision makers in the public, for-profit and nonprofit sectors.

To summarize, we recommend that the United States create a system that will assess workforce readiness. This will address the first dimension of student learning. Such a system could be created in a relatively short time period, in contrast to the other facets of this proposal that would take considerably more time and resources before they can be implemented as a truly comprehensive, nationwide system.

Domain-Specific Knowledge

The second major dimension of student learning that can be assessed in a comprehensive, nationwide system involves the establishment of agreement among experts in the respective fields as to what constitutes core knowledge/skills that should be expected of members of a given field. As indicated earlier, many professions (e.g., nursing, medicine, law, accounting) have certification programs in place, and these represent important initial components of an overall system.

To begin work on this dimension of student learning, we recommend convening a series of expert panels to address the issue of what constitutes adequate knowledge and preparation within broad major fields. Although there is a constant evolution of new, often interdisciplinary fields of inquiry within higher education, at the undergraduate level there is still some commonality of methods of inquiry and fundamental findings that characterize broad fields such as economics, chemistry and literature.

Because there is less consensus, compared with workplace readiness and general skills, as to what constitutes adequate basic knowledge within the academic disciplines, this second component of our recommendation would necessarily take longer to put in place. Our recommendation is that the next step be to charge the appropriate academic organizations with the task of articulating a set of expected learning outcomes for that field. Importantly, these learning outcomes would need to be specified at both the associate's and the bachelor's degree levels. As with the workforce readiness measures, there would need to be developed a comparable set of measures that could be used for pre- and post-measures, or input and output measures.

Soft Skills

For the third dimension of student learning, we make a somewhat different recommendation. Although there is growing recognition of the importance of these skills in higher education and the workforce, the present state of the art in assessing these skills is not adequate for supporting the institution of a nationwide set of standardized measures. As a result, we recommend that researchers continue to explore the development of these measures and that colleges be encouraged to incorporate those measures that they see as appropriate for their institutions into their overall assessment plans.

Although we are aware that many organizations and industries now regularly utilize a range of measures of soft skills in hiring and promotion decisions, we do not recommend that soft skills be included immediately in the comprehensive system of accountability for student learning in higher education. There are several important facts that lead to this aspect of our recommendation. First, if a set of soft-skill assessments were to be included in a nationwide system of higher education accountability, this would immediately make them high-stakes assessments. That is, if accreditation or resource allocation decisions were dependent on the results of assessments of soft skills, then one would need to consider the extent to which these assessments would be susceptible to short-term coaching effects, among other threats to their validity. At the present state of the art in assessing soft skills, the assessments are, unfortunately, susceptible to such undesirable coaching effects. There are, of course, a host of differences between assessment practices in academic vs. workplace settings. One of the fundamental differences that leads directly to our recommendation is that employers can and do select from a wide array of tools for assessing soft skills. In addition, they use these tools with far fewer people than we envision for postsecondary education settings. A nationwide higher education accountability system would have to rely upon a small number of well-validated measures, given to many more individuals.

Given the high-stakes nature of such academic assessments, coupled with the fact that it would be widely known which assessment tools are to be used in the academic arena, the currently available assessments would be less than ideal. We therefore recommend that further work be done to develop soft-skill assessment approaches that are not limited in these and other ways.

Student Engagement

Student engagement is an important aspect of the functioning of an institution, and the data from surveys such as the NSSE and CCSSE are surely important to faculty and leaders of colleges and universities. It is important, however, to keep in mind that measures of student engagement tell us about the *learning process*, but they do not measure *what students have learned*. Measures of student engagement thus provide leaders within the academy with valuable information that can be used *in conjunction with direct measures of student learning* to paint a more complete picture of how the institution is functioning, where there are strong points of institutional performance, and where there may be areas of concern.

In light of this, we recommend that academic leaders start or continue to collect data on student engagement. We do not, however, recommend that this become part of the proposed comprehensive system for assessing student learning. It will take considerable resources to institute systems to measure the two primary forms of student learning proposed here for immediate assessment — workforce readiness and domain-specific knowledge. We recommend that national efforts be directed at achieving these two more-proximate goals while researchers and others continue to address issues in high-stakes assessment of soft skills, and individual institutions determine the manner and extent to which they wish to examine student engagement.

Key Design Features of the Proposed Assessments

Several features of the recommended system would increase the overall value of this system to the various stakeholders in higher education. In this section we identify these features and consider some of the practical implications of our recommendations. Unless otherwise noted, these features pertain to both workforce readiness and domain knowledge. Because we are not advocating the inclusion of measures of soft skills or student engagement in this system at the present time, we will not address these two domains further in this paper.

Sampling and Modularization

As noted above, we recommend using sampling procedures to avoid having to test all students in higher education each year. A sampling procedure has the advantage of being cost-effective, leveraging existing statistical and psychometric approaches to yield data that are fair, useful and valid. Related to the sampling procedure recommendation, we further recommend that the measures be modularized such that each module can be completed within 50 minutes.

Many pragmatic implications and opportunities grow from a sampling and modularization approach. First, sampling, by definition, allows the development of measures that do not need to be completed by every student. Given the range of abilities to be assessed, testing all students would be extremely expensive and time-consuming.

Second, using a sampling procedure at the national level does not preclude an institution from oversampling (i.e., sampling more students than would be required to meet the needs of the comprehensive program). This gives institutional leaders the opportunity to extend the assessment plan in ways that are responsive to their institution's individual needs and resources.

Third, modularized assessments can be incorporated into regularly scheduled undergraduate courses and schedules. Depending upon how the testing is completed, it is possible to integrate

these assessments appropriately into target courses and to have the scores on the assessments considered in assigning grades. There is a wide range of options, but the important point is that this would allow individual institutions to address such issues as motivating students and identifying the best sampling procedures within the confines of their own cultures and resources.

Locally Developed Measures

It is essential that institutions be allowed to include additional items, questions, tasks and other elements to supplement the nationally developed and administered standardized tests. This feature allows an institution's faculty to retain control over those aspects of assessment that are important to that institution. It also gives increased institutional flexibility and allows for options such as groups of schools or programs deciding upon specific questions that they wish to address, and then developing and including measures that will provide answers to these questions.

Constructed Responses

Currently available technologies allow for the efficient and accurate scoring of constructed responses such as essays and short answers, and these technologies will certainly improve in the future.[7] In light of the many concerns voiced within the academy about exclusive reliance on multiple-choice question types, it is important that at least some of the measures used in this comprehensive assessment involve constructed responses.[8]

Pre- and Post-Learning Measures/Value Added

We recommend that appropriate measures be developed that will allow institutions to assess their students' levels of mastery at both the entrance and exit stages of their careers. Given the fact that this assessment system is intended to be applicable to both two- and four-year institutions, this means that there would actually be three levels of the assessment, corresponding to entering college, completing the associate's degree, and completing the bachelor's degree.

The availability of data from these three points within the education system will allow researchers to apply various types of value-added models (Braun, 2005).

Regular Data Collection

To track the progress of IHEs as they seek to improve student learning, it is critical that the types of student-learning data we are recommending be collected on a regular, ongoing basis. Because the approach advocated here involves a sampling procedure, conceivably, the learning data could be collected annually without great expense. If it is not feasible to collect data annually, then every other year may be adequate. What is important is that institutions are aware from the outset that this will not be a "one-off" exercise, or an exercise that can be completed once every 10 years.

Focus on Institutions

Although there should be other options for more extensive institutional data collection efforts, our recommendation is that student learning be sampled, rather than measured by testing every student. A corollary of this sampling approach relates to the appropriate conclusions that may be drawn from these data or the unit of analysis. This recommendation is explicitly intended to have institutions (in the case of workforce readiness) or departments/programs within institutions (in the case of domain knowledge) function as the primary unit of analysis. This has important implications for how the data from these assessments can be used.

First, for policymakers, academic leaders, boards of trustees and other stakeholders, having data on the performance of schools and programs would allow them to assess institutional progress over time and such issues as the impact of resource allocation decisions. For students

and their families, this focus would yield important data to allow cross-institution comparisons that would assist them in making decisions on choosing colleges that meet their needs. For employers, the focus on institutions would mean that although the employers have access to data on how well students *on average* from various schools are performing, it would not be possible to have information on how well individual students performed. (Again, however, this could be possible if the institution decided to test all students and report scores on individual students.)

Faculty Involvement

For this system of accountability for student learning to work as planned, it is essential that representatives from the faculty be integrally involved in the definition of the skills and abilities to be measured and in the development of the assessment tools. It is important to note that this does *not* mean that the faculty needs to be responsible for the creation of the measures, or that there be unanimous agreement among faculty as to what the exact skills and knowledge bases are that will be assessed.[9]

Comparability Across Institutions: Standardized Measures

Any institution that is currently accredited by one of the six regional accrediting associations within the United States is doing something — and in many cases many things — to assess student-learning outcomes. These many efforts reflect genuine concern among faculty for having appropriate processes within each institution to ensure that students have been exposed to an appropriate curriculum, that the students have been given clear notice of what is expected of them, and that faculty have appropriate grading policies in place. In addition, many institutions use a common set of outcome measures or procedures across their campuses to measure important skills, such as writing ability. These measures may be commercially available standardized tests or locally developed tests.

Our recommendations are not intended to replace these institutional initiatives, although is it conceivable that many institutions will see greater value in the set of measures being recommended here. Thus two additional, critical features of the measures we propose are that: (a) they are used at all institutions, thereby allowing comparisons across institutions and/or groups of peer institutions; and (b) the measures are assessing a common set of skills/abilities/knowledge bases. Together, these two features represent an important step forward in providing empirical data that would give stakeholders in higher education a method for studying the performance of U.S. higher education, and for making the value of postsecondary education clearer to all. Unless we adopt a common framework and set of measures, we will never be able to make the types of comparisons required by the national calls for accountability in higher education.

Summary of Key Design Features

The system that we are recommending would create an invaluable, and currently nonexistent, national resource. The resultant database for workforce readiness and domain knowledge would be useful to policymakers, students and families, employers, and researchers and other academics. The system is intended to be a mixture of standardization, in terms of the domains assessed and the measures used, and customization. Institutions, or groups of institutions, would be encouraged to add to the basic assessments in order to gain information relevant to their missions. By including an emphasis on constructed-response measures, we are taking advantage of improvements in technology, statistical techniques, and psychometric models to create assessments that move significantly beyond the types of multiple-choice, large-scale assessments that have been used in higher education in the past.[10]

Finally, the focus on institutions and programs within institutions is appropriate for a broad range of questions that are being asked by policymakers, legislators, employers and students. Academic institutions would be free to use the data from these assessments, along with other data (e.g., NSSE or CCSSE data on student engagement), to guide their efforts to improve undergraduate learning. Importantly, the data that would result from this comprehensive assessment could also be used within the context of the quality-improvement initiatives currently under way in the regional accrediting agencies (e.g., the *Quality Enhancement Plan* used by Commission on Colleges of the Southern Association of Colleges and Schools, and the *Academic Quality Improvement Program* used by the Higher Learning Commission of the North Central Association of Colleges and Schools). It would also allow for important benchmarking and peer-group comparisons in these processes, something currently not possible.

Implementing the New System: The Role of Accrediting Agencies

A nationwide system of accountability needs to be developed within the context of ongoing efforts to monitor and improve higher education. In his recent, sweeping review of undergraduate student learning, *Our Underachieving Colleges: A Candid Look at How Much Students Learn and Why They Should Be Learning More*, Derek Bok (2006) examines many of the issues regarding responsibility for ensuring student learning. An important dimension of Bok's analysis is his review of the organizational processes that have historically limited the extent to which colleges and universities have tended to focus real efforts on improving undergraduate learning.

Although Bok arrives at different conclusions than those presented here, he notes organizational impediments to efforts to improve undergraduate learning. Even a brief review of these impediments is beyond the scope of this paper, but we offer a suggestion that takes many of them into account. *We recommend that the six regional postsecondary accrediting agencies be charged with integrating a nationwide system of assessing student learning into their ongoing reviews of institutions of higher education.*

There are several reasons why these agencies are ideally positioned to meet this need. First, the regional accrediting agencies are already focused on reviewing the processes that institutions use to monitor various aspects of their functioning. Second, because they have both national and regional responsibilities, these organizations have the appropriate perspective for this effort. Third, because the organizations are charged with doing *regular* reviews, accountability for student learning will be an ongoing part of institutional operations.

Additional Themes in Higher Education Accountability

Our current recommendations deal with only one of the three themes identified by the National Forum convened by the National Center for Public Policy and Higher Education. In addition to the critically important theme of the performance of the college-educated population within a state, the forum also called for information on state literacy levels and the readiness of college graduates for advanced practice. Advanced practice here refers to "the proportion of the state's college graduates, from both two- and four-year institutions, who are ready for advanced practice in the form of vocational/professional licensure or graduate study." (Miller & Ewell, 2005, p. 9).

We agree that these two additional aspects of educational attainment are important, and we recommend that they receive continued research and policy-analysis attention. The approaches used by Miller and Ewell to assess these domains are reasonable, and we believe that they will be scalable to the national level. To assess the literacy levels within the state population, these investigators used the National Assessment of Adult Literacy (NAAL). This assessment

provides important benchmarking data on proficiency levels in prose literacy, document literacy, and quantitative literacy. The approach taken to assess the preparedness of state residents for advanced practice was to examine performance levels on three types of existing standardized examinations: licensure examinations, admissions examinations (e.g., MCAT, GRE), and teacher preparation and licensing measures. As noted above, admission measures are not taken by the entire college-going population as defined here. Further, NAAL is not appropriate for measuring the wide range of student learning that a nationwide system would need to address, either in breadth or depth of coverage of student learning.

We applaud the efforts to assess states in these two important areas. Going forward, it will be important to keep these efforts in mind as we consider the implementation of a nationwide system to assess student learning in workforce readiness and domain-specific skills and knowledge.

"Blue Sky": A Continuum of Possibilities and Next Steps

Education leaders, policymakers and other stakeholders are acutely aware of the need for highly innovative solutions that present the opportunity for significant change, even if the risk of failure is also strongly present. With this in mind, and based on the considerations outlined earlier, we recommend that an expert panel be convened to review the Assessment Framework Template (see Table 1). The expert panel would be charged with three activities: (a) reviewing the dimensions of learning and their subparts to reach consensus on the framework, (b) reviewing the completeness of the list of assessments[11], and (3) reviewing each assessment to determine whether it measures each of the agreed-upon dimensions of student learning for both two-year and four-year institutions. Once the table is populated, a picture of national postsecondary education would be available for use in accountability and improvement of postsecondary education.

Given the likelihood of a mixed early reaction to the general concept of postsecondary education assessments, an incremental approach to implementation may be appropriate for initial consideration. Here are several related issues for consideration:

- Regarding assessment development, the options range from having one organization develop and test the needed assessments to the clearly less desirable option (from the point of view of comparability and efficiency) of having each of the 4,071 institutions develop its own assessments.
- The outcomes associated with successful performance on the different dimensions of student learning could vary. For example, mastery of work-readiness skills could lead to a certificate, while performance on domain areas could be tied to a new valuation of the bachelor's degree.
- Performance indicators could be developed for individuals, institutions or both.
- The number of students taking the assessment could range from all students in higher education to a sample from each institution.
- The number of times that students take the assessments could range from one to multiple times.

Several key questions may guide the expert panel as it considers where on the different continua it wishes to place its marks:

- Should there be individual scores? Would this help future employers and graduate and professional schools know more about the inputs into their systems? How should this consideration be balanced with the cost savings of a sampling approach?
- Should there be institutional scores? Would an institutional score help both prospective students and their families have a more informed sense of what the educational experience will be like? What would an institutional score signal to employers and graduate and professional schools about their graduates?
- What should the rollout plan be for the new postsecondary education system? Should a demonstration program be conducted while plans for a longer-term nationwide system are developed?
- What are the desired types of analyses — pre-/post-test, individual growth models, value-added analyses? Each of these analyses has important data thresholds that need to be met.

Table 1

Assessment Framework Template

Assessments	Workforce Readiness and General Education Skills	Domain-Specific Knowledge	Soft Skills	Student Engagement
ACT Workkeys				
CAAP				
CCSSE				
CIRP				
CLA				
College Base				
Collegiate Readiness Survey				
Dublin Descriptors				
ETS GRE				
ETS GRE Subject Tests				
ETS Major Field Tests				
ETS MAPP				
Georgia Regents' Test				
Licensure Tests				
NAALS				
NSSE				
Oregon PASS				
UAP Field Tests				

References

Adelman, C. (2006). *The Toolbox revisited: Paths to degree completion from high school through college*. Washington, DC: U.S. Department of Education.

American Education Research Association. (2000). *High-stakes testing in pre k-12 education*. Retrieved February 13, 2006, from http://tinyurl.com/q6yb4.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). S*tandards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Astin, A. W. (1993). *What matters in college: Four critical years revisited*. San Francisco, CA: Jossey-Bass.

Bok, D. (2006). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton, NJ: Princeton University Press.

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: ETS.

Chronicle of Higher Education. (2005). *The 2005-6 Almanac: Proportion of colleges using various selection criteria for undergraduate admission, 2003-4*. Retrieved February 17, 2006, from http://tinyurl.com/erb7l.

Council for Aid to Education. (2005). *Collegiate Learning Assessment [CLA]: CLA in Context 2004-2005*. New York: Council for Aid to Education.

Joint Quality Initiative. (2004, March 23). *Shared "Dublin" descriptors for the bachelor's, master's and doctoral awards*. [Draft 1.31]. Retrieved February 13, 2006, from http://tinyurl.com/r4nvr.

LeMon, R. E. (2004). The changed social compact for public higher education: What do the public and lawmakers need? In *Proceedings of a National Symposium: A new compact for higher education: Accountability, deregulation, and institutional improvement*. Austin: The University of Texas System.

Malandra, G. H. (2005, December). *Creating a higher education accountability system: The Texas experience*. Commission Report presented to A National Dialogue: The Secretary of Education's Commission on the Future of Higher Education, Washington, DC: U.S. Department of Education. Retrieved February 28, 2006, from http://tinyurl.com/z3fxc.

McGrath, C. (2006, January 8). What every student should know: Even Harvard as it replaces the well-known CORE isn't quite sure. *New York Times, pp. 32-35.*

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Miller, M. A., & Ewell, P. T. (2005). *Measuring up on college-level learning.* Retrieved February 29, 2006, from National Center for Public Policy and Higher Education Web site: http://www.highereducation.org/reports/mu_learning/index.shtml.

National Assessment of Adult Literacy. (2006). *Demographics*. Retrieved February 28, 2006, from http://tinyurl.com/fkkeb.

National Center for Educational Statistics. (2002a). *Digest of education statistics 2004, chapter 3, postsecondary education, table 214*. Retrieved February 28, 2006, from http://tinyurl.com/negxr.

National Center for Educational Statistics. (2002b). *Digest of education statistics 2004, chapter 3, postsecondary education, table 175*. Retrieved February 28, 2006, from http://tinyurl.com/qje9c.

National Center for Educational Statistics. (2002c). *Digest of education statistics 2004, chapter 3, postsecondary education, table 255*. Retrieved February 28, 2006, from http://tinyurl.com/zako8.

National Survey of Student Engagement. (2005a). *The college student report*. Retrieved February 22, 2006, from http://tinyurl.com/pnfuj.

National Survey of Student Engagement. (2005b). *NSSE 2005 annual report: Exploring different dimensions of student engagement*. Retrieved February 22, 2006, from http://nsse.iub.edu/NSSE_2005_Annual_Report/index.cfm.

Parsad, B., & Lewis, L. (2003). *Remedial education at degree-granting postsecondary institutions in fall 2000*. (NCES 2004-010). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Peter, K., & Forrest Cataldi, E. (2005). *The road less traveled? Students who enroll in multiple institutions* (NCES 2005–157). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Romano, L. (2005, December 25). Literacy of college dropouts is on decline. *Washington Post*. Retrieved February 22, 2006, from http://nsse.iub.edu/html/quick_facts.cfm.

Toppo, J. (2006, January 31). Spellings 'living' her job; Education chief says her kids help her relate [interview with Secretary of Education Margaret Spellings]. *USA Today*, p. 8D.

ENDNOTES

[1] The Carnegie Classification system, a well-known system for categorizing institutions into similar types, has recently acknowledged that there are many ways to categorize IHEs, depending upon the goal of the classification. Carnegie recently introduced a system that includes five dimensions: Undergraduate Instructional Program, Graduate Instructional Program, Enrollment Profile, Undergraduate Profile, and Size and Setting. See http://www.carnegiefoundation.org/classifications/.

[2] Institutions review different academic credentials when admitting first-year students and transfer students to their institutions.

[3] We focus here on student learning rather than teaching to highlight that the real focus needs to be directly on learning no matter how it occurs. Teaching is one way of allowing students to learn, but there are many other ways (e.g., creating opportunities for internship experiences, study abroad, etc.) to create a learning experience that does not rely upon traditional teacher-student roles.

[4] Advanced standing refers to students who have acquired academic credits that allow them to start college with earned credits.

[5] See, for example, Bok (2005), Chapter 12, for a discussion of the importance of the efforts of individual institutions.

[6] Even this type of controlled experiment is not completely adequate for drawing conclusions about the effects of students taking courses at University X. There are many other differences between the experiences of University X students vs. the control group (e.g., students at University X come in contact with many other college students whereas individuals in the control condition may not have had this type of experiential learning).

[7] As one example of these rapidly developing technologies, ETS has developed the Information Communication Technology (ICT) Literacy assessment, which uses a series of simulated tasks (e.g., collecting information from simulated Web sites to develop a presentation) to assess literacy in the digital age. This approach can easily be expanded to other skill and knowledge domains.

[8] It is also important to note that one result of the utilization of automated scoring of constructed responses may be that faculty come to appreciate the power of these newer technologies and begin including them in their toolbox of teaching tools. This could help, for example, with the need to provide students with more practice in writing and revising.

[9] It is well known in the academy that it is often difficult to get 100% faculty buy-in on any standardized assessment, and the proposed system will very likely be challenged by some faculty. However, given the fact that faculty own the curriculum at many institutions, and the respect that academics have for faculty control, it is essential that faculty play a role in the development of the assessments of workforce readiness and domain-specific knowledge. One model that could be used to address this need is the model that has been used by the GRE Program in creating the GRE Subject Tests. A faculty committee of experts within the discipline works with test development specialists to create the test specifications (e.g., what content areas within the field will be assessed in the test), review test questions, and make certain that the test remains current with the discipline as it is taught across the country. This model leverages the expertise of faculty and their appropriate role in defining important content knowledge and also the expertise of test development specialists to create psychometrically valid assessments.

[10] We expect that the final set of measures will involve a combination of various item types, including multiple-choice and constructed response. Both classes of item types have appropriate uses and should be considered.

[11] A public notice could be issued to allow for the public to provide examples of other assessments for consideration.

Listening.
Learning.
Leading.

Copyright © 2006 by Educational Testing Service. All rights reserved.
Educational Testing Service, ETS, and the ETS logos are registered trademarks of Educational Testing Service.