The Case for Generic Skills and Performance Assessment in the United States and International Settings

Roger Benjamin • Stephen Klein • Jeffrey Steedle
Doris Zahner • Scott Elliot • Julie Patterson

NEW YORK, NEW YORK: CAE-OCCASIONAL PAPER 1, MAY, 2012

forward

A testing program like the Collegiate Learning Assessment (CLA) is constantly evolving. CAE (Council for Aid to Education) learns from the hundreds of colleges and universities that test with either the CLA or, its high school counterpart, the College and Work Readiness Assessment (CWRA) about what works well, what can be improved, and what new ideas are to be considered for aligning assessment with teaching and learning. We also benefit from the criticisms, comments, and suggestions made through public communications. Working constructively with these criticisms, along with extensive in-house research, has improved the reliability and validity of the CLA. These improvements have led to new concepts for enhancing our existing protocol, such as through CLA Education, which focuses on the teaching and learning implications of the CLA performance assessments. Additionally, new ideas for administering the performance assessments in way that will provide formative and useable results at the student level are being formulated, piloted and—in the near future—implemented.

Recently, we at CAE pulled together the full range of perspectives on assessing higher-order skills (or as they are called internationally, generic skills). These perspectives include studies on: the reliability and validity of the CLA performance assessments (conducted both by both third parties and CAE measurement scientists), the place of assessment in the knowledge economy, and on the major goals for student-learning assessment today.It's been a valuable exercise and I believe it will be useful to anyone interested in assessment. United States postsecondary education is entering a turbulent period. Faculty and administrators struggle to educate students in the face of rising costs, declining resources, the challenge of education technology, and the need to figure out how to re-allocate resources while improving undergraduate education at the same time. These are not small problems. We will note here some bold experiments that give the promise of solutions to the problems noted. Here, then, is a short monograph divided into two parts. Part I is the Case for Generic Skills and Performance Assessment in the United States and International Settings.[1] It attempts to note third party studies of the CLA, while referencing our responses to them. Part II places this general argument--along with studies of reliability and validity--in a larger policy context. Using a political economy framework, I suggest why the general assessment community should expect the role of evidence-based decision making in postsecondary education to grow, including the provision for a more central role for assessment of student learning. I also review the main reasons for not engaging in any standardized assessment.

We hope this monograph provides readers with useful information about both assessment and larger education policy issues. It is a critical subject for all of us.

Roger Benjamin
President, CAE

[2] We refer to generic skills because we had the international audience associated with the Assessment of Higher Education Learning Outcomes project (AHELO) in mind as well as the United States audience when writing this paper.

INTRODUCTION

Educational institutions across the world are being challenged to improve instruction so that tomorrow's workforce will have the knowledge and skills necessary to meet the demands of modern careers, while contributing to the global economy. Indeed, a college education has never been more necessary for productive participation in society. Employers now seek individuals who are able to think critically and communicate effectively in order to meet the requirements of the new knowledge economy (Hart Research Associates, 2006; Levy & Murname, 2004). Therefore, the skills taught in higher education are changing; less emphasis is placed on content-specific knowledge and more is placed on general higher-order skills, such as: analytic reasoning and evaluation, problem solving, and written communication.

Any rigorous improvement project requires constant evaluation in order to measure progress toward goals. Consequently, there is a clear need for standardized assessments that measure general higher-order skills, such as the Collegiate Learning Assessment (CLA). Performance assessments like the CLA evaluate not only whether students are learning the higher-order skills required of today's workforce, they also spur educational advances in pedagogy. The CLA presents students with scenarios that are representative of the types of problems they will encounter in the real world and asks them to generate solutions to these problems. Unlike multiple-choice questions where students need only to identify the correct answer—limiting the capacity of those questions to measure students' critical thinking skills—an open-ended assessment such as the CLA is able to measure how well students formulate hypotheses, recognize fallacious reasoning, and identify implicit and possibly incorrect assumptions. Only open-ended tasks can authentically capture this type of critical thinking, as well as the ability to organize and present ideas in a coherent argument.

Because higher-order skills are so critical to national productivity, CAE supports the concept of the Generic Skills Strand of the Assessment of Higher Education Learning Outcomes (AHELO) project. From a methodological perspective, CAE believes that the Generic Skills Strand is the only solid basis for cross-national benchmarks that are inclusive of all academic disciplines. Of course, knowledge and skills specific to academic disciplines are important, but there is a multitude of disciplines, each potentially differing across national contexts and evolving over time. This makes it impractical to establish broad, cross-national benchmarks based on achievement in academic disciplines. The approach of the Generic Skills Strand is to establish benchmarks for student achievement of essential higher-order skills that cut across national contexts and academic disciplines. The development of students' generic skills is central to the missions of modern postsecondary institutions because of growing recognition that these skills fuel innovation and economic growth (Levy & Murname, 2004). The first section of this paper provides a rationale for focusing on generic skills in society and describes how these skills are operationalized in the development of performance tasks for the CLA. The next section describes the CLA, summarizes a decade's worth of validity research pertaining to the use of the CLA in postsecondary institutional assessment programs, and addresses common concerns and critiques related to the CLA. The final section presents a summary of the case for measuring generic skills.

RATIONALE FOR FOCUSING ON GENERIC SKILLS

Political and economic leaders everywhere understand that workforce skill level is what determines economic performance. This understanding has led policy analysts to view education policy as being equally important as other critical policy fields, such as: healthcare, national security, international trade, and the environment. In other words, education policy is now viewed as one of the top societal or governmental priorities. The initial credit here goes to Gary Becker and his colleagues in the economics department at the University of Chicago, who developed the human capital school of labor economics. They leveraged the methodological rigor of contemporary economics to formulate the principles of human capital over forty years ago (Becker, 1964; Heckman & Krueger, 2003).

Their achievements have been accepted at the highest levels of the academy, including recognition of several members of the human capital school by the Nobel Committee.

These scholars defined human capital as the stock of knowledge and skills present in a nation's population. Such capital accrues through education, training, and experience. As the field matured, economists began to mine its implications for education, which is the formal venue for human capital development. Analysis of the returns on the amount of education achieved has become an important academic pursuit in economics, public policy, and education. This body of research suggests that education must focus on the stock of knowledge and skills required by a society which today most highly values the ability to access and structure information and apply it to solve new problems.

Recent theories of learning that reflect the change in emphasis from specific content domains to a focus on higher-order skills are redefining the concept of knowledge. Herbert Simon (1996) argues that the meaning of "knowing" has changed from being able to recall information to being able to find and use information. Branford, Brown, and Cocking (2000) note that the "…sheer magnitude of human knowledge renders its coverage by education an impossibility; rather, the goal is conceived as helping students develop the intellectual tools and learning strategies needed to acquire the knowledge to think productively."

The logical extension for some is to say that education should be more explicitly vocational, but this is not the point. As the world economy has evolved from the industrial era to the knowledge economy, it has become increasingly dependent on a workforce that can generate knowledge that can be a foundation for economic prosperity. Knowledge generation requires strong skills in analytic reasoning, problem solving, and writing—referred to as core generic skills. Thus, education must prepare students for productive participation in the economy and society, and increasingly this means teaching generic skills and measuring progress toward desired achievement levels.

MEASURING GENERIC SKILLS

Increasing recognition of the essential role of generic skills in the knowledge economy portends significant changes in teaching and learning. This is reflected in the educational reform movement now underway and assisted by education technology. Although this reform is present in elementary and secondary education, most advances have occurred in postsecondary or tertiary education in Europe and the United States. The reform movement can be characterized along three dimensions:

- Shifting from the long-standing lecture format to a student-centered approach emphasizing students' active class participation and development of analytic writing skills.
- Changing the balance of curricular and textbook focus from its current emphasis on content to case studies and problem-based materials requiring students to apply what they know to novel situations.
- Changing assessment instruments from multiple-choice tests that are best used for benchmarking the level of content absorbed by students to open-ended assessments that are aligned with numerous goals of the reform initiative.

Although significant advances have been made on the first two dimensions of this education reform movement, assessment has lagged behind. As schools and colleges focus increasingly on developing generic skills in their students, assessment tools need to evolve to measure how well students are learning—and institutions are teaching—such skills.

Multiple-choice and short-answer tests remain the dominant testing regime, not only for facts, but also for generic skills. In the United States, they are used overwhelmingly by the Educational Testing Service (ETS), ACT, and the College Board. As a result, in postsecondary education and elsewhere, the testing regime is not assessing the most critical skills required of students in the workplace and—just as importantly—is not supporting the other two dimensions of

reform. We believe the promise of educational reform developing in today's knowledge economy cannot be achieved without employing open-ended, performance-based assessments, not only in postsecondary education, but in primary and secondary education as well.

As an illustration of this point, consider two tests of critical thinking: one multiple-choice and the other a performance assessment. To measure students' understanding of correlations and causality, the multiple-choice test requires students to select an answer from a list of four or five provided options. In the performance assessment, students are presented with a research report in which the author incorrectly concludes that there is a causal relationship between the two variables due to a strong correlation between them. The student must evaluate this information and determine how that information does or does not support possible solutions to a real-world problem. The cognitive processes involved in responding to these two assessments are fundamentally different. Recognizing the correct answer from a finite list of possibilities differs greatly from asking students to generate a critique and explain it clearly. In the latter approach, the student must not only recognize the fallacious reasoning but must also understand how the concepts are confused and explain why the argument fails. This level of fidelity to real-world experience is often viewed as a major advantage of performance assessments over multiple-choice tests. Additionally, performance assessments measure students' written communication skills and their ability to craft an argument and refute counterarguments with relevant and reliable information. Multiple-choice items that assess writing generally measure a student's ability to correctly identify proper use of vocabulary and grammar.

Another important advantage of performance assessments is that they are seen as tests worth teaching to. The practice of "teaching to the test" is generally frowned upon when referring to traditional multiple-choice and short-answer assessments, and there is ample evidence that this practice occurs, especially when educators are held accountable for their students' test performance. However, "teaching to the test" for performance assessments should be encouraged. That is, class time spent preparing students to apply knowledge and skills to complex, real-world problems is time well spent. If performance assessments are integrated into accountability systems, this has the potential to positively impact classroom practice by encouraging teachers to foster the development of competencies in generic skills. This effect has yet to be established, so it would be worthwhile to investigate whether the introduction of performance assessment for accountability purposes has the desired effect on teaching and learning. One potential barrier to investigate is the perceived level of effort required to use performance assessments regularly in the classroom.

In addition to negative effects on pedagogy, a critical shortcoming of today's principal educational assessment regime is that it pays little attention to how much a school or college contributes to developing the competencies students will need after graduation. For instance, the outcomes that are typically looked at by higher-education accreditation teams, such as a college's retention and graduation rates and the percentage of its faculty in tenured positions, say nothing about how well the school fosters the development of its students' analytic reasoning, problem solving, and communication skills. This situation is unfortunate because the ways in which institutions are evaluated significantly affects institutional priorities. If institutions were held accountable for student achievement, they would likely direct greater institutional resources and effort toward improving teaching and learning.

Compounding the challenges of implementing performance assessments is the fact that the development and measurement of performance assessments are rarely taught in schools of education or within the social sciences. Consequently, textbooks on assessment devote very little attention to this topic. The main focus of educational assessment courses and textbooks is item construction and analysis for multiple-choice tests. When performance assessment is taught in these programs, the focus is often on the development of performance tasks for professional licensure or certification purposes. For example, airline pilots are assessed on how competently they handle simulations of a range of realistic problems they may face. Similarly, mocked-up cases that require diagnosis by those studying to become medical doctors, veterinarians, or lawyers also widely use performance assessments (Heinrichs et al., 2007; Klein, 1996). There is hardly any attention devoted toward performance assessment in primary, secondary, or postsecondary education.

All these conditions point to the need to support advances in performance assessment, particularly in the field of education. If the human capital school demonstrates the importance of education, the implications of the knowledge economy and recent theories of learning place the focus on improving the higher-order skills of the next generation of students. These developments create an urgent need to generate and implement a testing paradigm that measures and simulates these skills. Because of its broad reach, AHELO is an excellent opportunity to demonstrate the effective use of performance assessment to measure generic skills.

WHAT THE CLA MEASURES

The CLA skills can best be understood by situating them in a cognitive framework. The framework views the range of potential outcomes as a continuum ranging from domain-specific knowledge to general ability, or *G* (Spearman, 1904).While the framework may be an oversimplification, it offers a basis for understanding the CLA and what it measures. At the top of the hierarchy are theories of intelligence, with Spearman (1904) at one extreme postulating a single undifferentiated general intelligence and Guilford (1967)and Gardner (2006)at the other end of the spectrum postulating multiple abilities and different independent intelligences.

The CLA is based on the belief, supported by research, that learning is highly situated and context-bound. However, through practice within a particular subject area, learned knowledge becomes sufficiently generalized to enable it to transfer to the realm of enhanced reasoning, problem solving, and decision-making skills that can be demonstrated across content domains. These broad abilities can be understood in relation to other major skills and abilities in the cognitive framework. The critical thinking skills measured by the CLA are broad abilities that are learned and applicable over an array of domains. The CLA does not measure the general reasoning abilities generally thought of as intelligence or *G*, nor is the CLA measuring the domain-specific skills limited to one or a few disciplines.

Critical Thinking

While there are many desirable outcomes of college education, there is widespread agreement that critical thinking skills are among the most important. As Derek Bok (2005), former president of Harvard University, states, "with all the controversy over the college curriculum, it is impressive to find faculty members agreeing almost unanimously that teaching students to think critically is the principle aim of undergraduate education" (p. 109). Critical thinking skills are longstanding desired outcomes of education (Dewey, 1910; Educational Policies Commission, 1961), and in modern day, they are seen as essential for accessing and analyzing the information needed to address the complex, non-routine challenges facing workers in the 21st century (The New Commission on the Skills of the American Workforce, 2006; The Secretary's Commission On Achieving Necessary Skills, 1991). In recognition of the central role that critical thinking plays in the information age, leaders in higher education, business, and government stress that such higher-order skills must be assessed at the college level (Business-Higher Education Forum, 2004; Silva, 2008; State Higher Education Executive Officers, 2005; U.S. Department of Education, 2006).

Despite variation in definitions of critical thinking, there is significant agreement on its core components. The American Philosophical Association's (1990) definition, which reflects the consensus of 200 policy makers, employers, and professors, describes critical thinking as: "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference as well as explanation of the evidential, conceptual and methodological considerations on which a judgment is based" (p. 2). Along these lines, Pascarella and Terenzini (2005) offer an operational definition of critical thinking largely based on the work of Erwin (2000):

> Most attempts to define and measure critical thinking operationally focus on an individual's capability to do some or all of the following: identify central issues and assumptions in an argument, recognize important relationships, make correct references from the data, deduce conclusions from information or data provided, interpret whether conclusions are warranted based on given data, evaluate evidence of authority, make self-corrections, and solve problems (p. 156).

Bok's (2006) definition of critical thinking captures similar qualities:

> The ability to think critically—ask pertinent questions, recognize and define problems, identify arguments on all sides of an issue, search for and use relevant data and arrive in the end at carefully reasoned judgments—is the indispensable means of making effective use of information and knowledge (p. 109).

The aspects of critical thinking measured by the CLA are well aligned with the definitions of critical thinking provided above. Note that critical thinking may be defined very broadly, so we include analytic reasoning and problem solving in this construct definition (and in other CLA documentation) in order to expand upon the critical thinking skills measured by the CLA and to denote the range of those skills. Students are judged on critical thinking skills such as analytic reasoning and problem solving during the scoring process, which captures qualities exhibited in student work such as evaluating the reliability and relevance of evidence, identifying logical flaws and holes in the arguments of others, analyzing and synthesizing data from a variety of sources, drawing valid conclusions and supporting them with evidence and examples, and addressing opposing viewpoints. Students obtain higher CLA scores by attending to specific items in a task (e.g., accurately interpreting a graph or identifying a statement as untenable, given other information the examinee receives) and by applying the skills described above generally (e.g., overall strength of support for arguments).

Writing

In addition to critical thinking skills, colleges are expected to teach "top notch writing and speaking skills" (Immerwahr, 2000, p. 10). This derives from recognition that, in many professions, the ability to communicate ideas effectively and articulate problem-solving processes is an important and highly-valued skill. In response to CLA prompts, students generate text that describes an analysis of a problem, provides evidence and examples to support a position, explains weaknesses in the arguments of others, and proposes a course of action. CLA scoring rubrics capture how well students write in a style that is well-organized, persuasive, and free from grammatical errors.

THE CLA APPROACH TO MEASUREMENT OF GENERIC SKILLS

Unlike most traditional learning assessments, that grow out of an empiricist philosophy and a psychometric/behavioral tradition, the CLA employs a criterion sampling/competency measurement approach (Shavelson, 2008). Traditional learning assessments take everyday complex tasks, divide them into components, create measures for each individual component (most often using multiple-choice questions), collect scores on each measure, and sum those scores to describe examinee performance. The problem arises when one tries to generalize back to the broader domain presented by those everyday tasks; the deconstructed measures may bear little resemblance to the complexity of the everyday tasks upon which they were based.

In contrast, the CLA is based on a combination of

rationalist and socio-historical philosophies in the cognitive constructivist and situated-in-context traditions (Shavelson, 2008). The criterion sampling approach employed by the CLA assumes that the whole is greater than the sum of its parts and that complex tasks require an integration of abilities that cannot be measured when deconstructed into individual components. The criterion-sampling approach is based on a simple principle: if you want to know what a person knows and can do, sample tasks from the domain in which that person is to act, observe his or her performance, and infer competence and learning (Shavelson, 2008). In short, the CLA samples tasks from "real-world" domains; the samples are holistic, real-world tasks drawn from life experiences. The samples require constructed responses (not selected) and elicit complex critical thinking, analytic reasoning, and problem solving skills.

This approach underlies the development of the CLA generic skills measure. Complex intellectual skills were identified; observable performances in the form of performance tasks were created and standardized, ensuring fidelity with real world criteria. The efficacy of this model is determined, in part, based on the interpretability of the inferences drawn from the individual's behaviour on a sample of tasks to what his or her behaviour would be on the larger universe of tasks (Shavelson, 2011). Both qualitative and quantitative evidence can be brought to bear here. Reliability and evidence of face, concurrent, and predictive validity offer support for those inferences.

THE CLA PROGRAM

The CLA represents a paradigm shift in testing and is a good example of how performance assessment can be used effectively. Unlike multiple-choice or short-answer tests, the CLA employs only performance tasks, which are concrete exercises that require students to apply a wide range of higher-order thinking and communication skills to solve a complex problem. In these tasks, students are allotted 90 minutes to examine a set of documents related to a real-world problem and write responses to explain their analysis of the documents and propose a solution to the problem at hand. The documents, which contain a mix of dependable and questionable information, appear as newspaper articles, research abstracts, emails, web pages, transcripts, graphics, maps, and other forms of written and visual media. CLA Performance Tasks are presented in a variety of contexts, including the arts, social sciences, natural sciences, business, education, political science, and other fields. However, no prior subject knowledge is required. Students use their analytical reasoning, problem solving, and writing skills to answer open-ended questions that are not framed to elicit "right" or "wrong" answers. Rather, students are asked to compose written responses requiring them to integrate information from the different provided documents and support their decisions with relevant facts and ideas.

There are a number of distinctive and noteworthy characteristics of the CLA:

- OPEN-ENDED PROBLEM SOLVING. In contrast to the typical multiple-choice and short-response items, the CLA relies on open-ended, realistic problems that are engaging and viewed as authentic by both students and faculty (Hardison & Vilamovska, 2008). Additionally, the CLA Performance Tasks are constructed to meet the highest standards of reliability and validity (Klein, Benjamin, Shavelson, & Bolus, 2007; Klein, Kuh, Chun, Hamilton, & Shavelson, 2005; Klein et al., 2009; Klein, Shavelson, & Benjamin, 2007).

- STAKES. Like most international comparative assessments (e.g., TIMSS and PISA), the CLA is a measure of typical (rather than maximum) performance. Consequently, given the CLA's track record of evaluating and measuring learning in higher education, it is an appropriate measure for AHELO Generic Skills Strand. Experience has taught assessment researchers that a low-stakes approach can provide valid assessment that avoids the negative incentives sometimes associated with high-stakes approaches.

- MATRIX SAMPLING. Most testing programs require that all students complete the same (or equivalent) long assessment instrument(s). These measures need to be long when (1) the test has to cover a broad content domain and/or (2) the results are used to make decisions about individual test takers, such as for college admissions or professional licensing. However, it is counter-productive to require that all examinees answer the same questions when the results are used to draw conclusions about groups of examinees, such as whether the students at one school (state, province or country) are performing at a higher level than those in another. That is why the United State's National Assessment of Educational Progress (NAEP) program assigns students randomly to questions. CAE also uses a matrix sampling approach. Specifically, it requires a student answer only a portion of the entire bank of questions that could be asked. This significantly reduces the testing burden on an individual student and allows for much better content coverage. Furthermore, there are enough students answering each question to ensure adequate content coverage and score reliability. This is why CAE recommends testing a random sample of at least 100 students per institution.

- UNIT OF ANALYSIS. CAE considers the institution, rather than the individual student, as the initial unit of analysis. Additionally, while higher-order skills measured by the CLA are not a function of any one particular course or even a specific academic major, some institutions are using the CLA to assess differences between programs. Nonetheless, the primary goal of the CLA is to be sensitive to an important portion of the summative effects of the courses and educational experiences students encounter over the course of an entire undergraduate program. This point is one of the most innovative features of the CLA: it assesses the institution, rather than the individual student.

- BENCHMARKING. The great majority of standardized assessments do not document the level of proficiency of their entering students. Thus, it is impossible to gauge how much improvement an institution itself contributes to the growth in student learning. The CLA controls for the competencies that students bring to the college, and results are reported in terms of "value-added" (i.e., how much value an institution adds to students over the period of time they are at the institution) and other indices. Research shows that CLA value-added scores are sufficiently reliable to make inferences about student learning relative to other institutions (Klein, Benjamin, et al., 2007; Klein, et al., 2005; Steedle, 2011 online first).

- VALUE-ADDED. An institution's CLA value-added score gives faculty and administrators a benchmark of where their institution stands relative to other institutions admitting students with similar entering academic ability. There is significant variation between similarly situated institutions along this value-added continuum. In other words, there are very large differences in CLA value-added scores among institutions that accept students with similar entering academic ability. This means there is a large canvas for studying best practices in the institutions that perform better than the equation predicts as opposed to those that perform worse. There is also ample opportunity for those institutions that perform less well than predicted to improve upon their contribution to their students' education.

- REPORTING. In reports to the institution, an institution's CLA value-added score is presented to provide an indicator of the growth in skills measured by the CLA relative to similarly selective institutions. In addition, absolute score levels are provided to show where an institution falls in the overall distribution before controlling for entering academic ability. The CLA results for each participating institution are sent only to that institution, but state politicians and other stakeholders occasionally require public reporting of some kind. Some institutions also share results publicly with prospective students as part of the Voluntary System of Accountability (McPherson & Shulenburger,

2006), a network of public, four-year colleges and universities who use a common web template for presenting information about institutional characteristics as well as student experiences and outcomes.

- INTERNET DELIVERY. An important feature of the CLA is its use of a secure Internet browser for administration and delivery of the assessment. The Internet has provided two important benefits for the CLA. First, the Internet-based delivery platform makes it possible to increase the complexity and richness of the performance assessments created. The performance assessments are comprised of a considerable number of pertinent documents which include tables, figures, and graphs. The Internet makes it possible to present and organize the information on the documents without overwhelming the students. Secondly, delivering the CLA over the Internet significantly reduces cost and frequency of errors related to test administration, scoring, and reporting. The CLA would not exist without the Internet.

PSYCHOMETRIC PROPERTIES OF THE CLA

There are several studies that speak to the reliability of CLA scores and to the validity of CLA score interpretations. Some of the key studies are highlighted below. A more comprehensive list of studies is given in Appendix A and the list of references.

Reliability

In institutional assessment programs, reliability is achieved when test results are consistent across different samples of students drawn from the same population. Here, the focus is on the reliability of aggregate institutional results rather than those of individual test takers. When the institution is the unit of analysis, the CLA's reliability is approximately 0.90 (Klein, Benjamin, et al., 2007). This indicates that the relative standings of institutions would be highly consistent if testing was repeated with different samples of students. Moreover, an institution's CLA value-added score has a reliability of approximately 0.75 (Steedle, 2011 online first).

Validity

Construct validity refers to the degree to which test scores can be interpreted as indicators of whatever skill (i.e., construct) the test purports to measure. While gathering validity evidence in any testing program is an ongoing activity, a substantial amount of validity research has already been done by both CAE researchers and independent third parties. Some of these studies are summarized below:

Face validity. An assessment is said to have face validity when it appears to measure what it claims to measure. In order for the CLA to have face validity, CLA tasks must emulate the critical thinking and writing challenges that students will face outside the classroom. These characteristics of the CLA were vetted by a sample of 41 college professors selected to be representative of faculty from a wide range of institutions (Hardison & Vilamovska, 2008). After an in-depth review of CLA Performance Tasks and reading a range of student responses, these professors completed a survey on their perceptions of the CLA Performance Tasks. As shown in Figure 1, results indicate that the professors considered the Performance Tasks to be good assessments of critical thinking, writing, problem solving, and decision making. For example, using a rating scale of 1 – 5, professors felt that the CLA measures what it intends to measure (Mean 4.14, SD 0.46); it measures important skills that college graduates should possess (Mean 4.70, SD 0.53); students need good critical-thinking skills to do well on the task (Mean 4.60, SD 0.46); and students who do well on the task would also perform well in a job requiring good

written communication (Mean 4.20, SD 0.83) or decision-making (Mean 4.10, SD 0.70). Respondents also agreed, after viewing the tasks, that college seniors should perform better on this task than college freshman (Mean 4.70, SD 0.48).

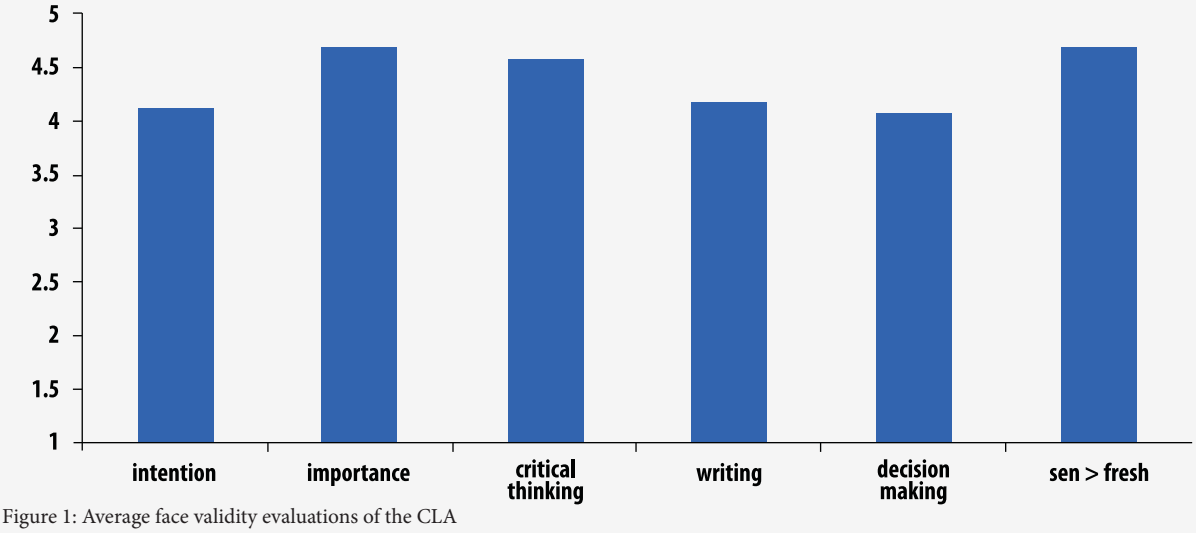

Figure 1: Average face validity evaluations of the CLA

Concurrent validity. Concurrent validity is commonly evaluated by examining the pattern of correlations between a test and other tests of similar and different skills (Campbell, 1959). For example, if the CLA measures critical thinking skills, then it should be highly (positively) correlated with other tasks that measure critical thinking. In the fall semester of 2008, CAE collaborated in a validity study with ACT and ETS to investigate the validity of the CLA, ACT's Collegiate Assessment of Academic Proficiency (CAAP) and ETS's Measure of Academic Proficiency and Progress (MAPP—currently known as the ETS Proficiency Profile) (Klein, Liu, et al., 2009). Results from the study show that for critical thinking, the CLA has a strong positive correlation with other tasks that measure critical thinking. The correlations at the institutional level between CLA scores and the critical thinking tests for MAPP and CAAP were .83 and .79, respectively. This evidence is consistent with the notion that the CLA measures critical thinking skills. Additional studies have also corroborated these results by showing that the CLA correlated highly with other measures of critical thinking (Carini, Kuh & Klein, 2006; Klein, et al., 2005).

In this context it is important to note that a moderate to high correlation between open-ended and multiple-choice test scores does not mean these measures assess the same construct. First, how one would prepare for a multiple-choice test is different than how one would prepare for an essay test. Secondly, a high correlation between the scores on a general-skills measure (such as the SAT or ACT) earned in high school and grades earned in a college-level organic chemistry course does not mean that high school seniors with high verbal and quantitative admission test scores know anything about organic chemistry.

Predictive Validity. The predictive validity of an assessment refers to how well a test score predicts some future criterion that is conceptually connected to the skills measured by the test. Traditionally, indicators of college readiness such as high school grade point average (HSGPA) and college entrance exam scores (SAT or ACT) are used to predict academic success in college as measured by college GPA. Results from a study using the CLA as a replacement for or supplement to college entrance exam scores showed that the most accurate prediction of students' senior-year GPA was achieved using the combination of SATs and the CLA scores (Zahner, Ramsaran, & Steedle, 2012). These results indicate that the CLA scores may capture knowledge and abilities that are different from content-based college entrance exams such as the SAT and ACT and underscore the apparent value of open-ended per-

formance assessments as evidence of college readiness and therefore as predictors of college success. Recent findings from a large multi-college longitudinal study found that students who perform well on the CLA as college seniors tend to have better post-graduate outcomes such as securing employment and having less credit card debt (Arum, Cho, Kim, & Roksa, 2012).

COMMON CLA CONCERNS AND CRITIQUES

Any new testing program that challenges the status quo and is used widely is bound to receive public and professional scrutiny, as well as generate criticism. This is especially so if it has or may have consequences for students, faculty members, and their schools. The CLA is not an exception. This section addresses the most common concerns that have been raised about the CLA that are relevant to AHELO (Appendix A summarizes the critiques of the CLA and CAE's responses to each critique).

Critique: If multiple-choice test scores are correlated with performance assessment scores, they provide the same information about student abilities.
A high correlation between two tests—for example, a multiple-choice critical thinking test and a CLA Performance Task—indicates that the relative standings of examinees on the two tests are similar. A high correlation does not necessarily mean that the two tests are providing the same information about student abilities. Indeed, it is common to find high correlations between obviously different constructs. For example, in the aforementioned validity study, the school average ETS Proficiency Profile Math and Writing test scores correlated .92. Put simply, a high correlation between two tests is consistent with the idea that they measure the same construct, but it does not prove that they measure the same construct (Steedle, Kugelmass, & Nemeth, 2010). Multiple-choice questions cannot adequately assess students' ability to use their analytic reasoning and problem solving skills to identify important strengths and weaknesses of arguments made by others, present a coherent, succinct, and well-organized discussion of the issues, and independently generate a solution to a real-world problem.

Critique: Performance assessments are too costly to administer and score.
Performance assessment is increasingly being implemented in large-scale testing programs because it is recognized as being more valid than multiple-choice testing. For instance, the new K-12 assessment programs being widely adopted in the United States are all focusing on the use of performance assessments. As a result, all of the US testing companies, including CAE, have made major commitments to building their capacity to develop and deliver performance assessments.

With this increase in demand, innovative approaches are being employed to address cost issues. Internet test delivery and reporting have been central to making performance assessment affordable. Computer-assisted scoring is also serving to reduce costs. Training human scorers, actual scoring, and maintaining scorer calibration account for a significant portion of the cost of performance assessments. Initially, human scorers need to be trained on how to score the student responses for performance tasks. However, once a sufficient number of responses have been collected, automated scoring in all languages is available. This lowers the cost of scoring Performance Tasks substantially, and the inter-rater consistency between two humans and between the computer and a human is comparable (Elliot, 2011).

Critique: Results do not generalize
Critics of the CLA make claims that sampling only 100 students will lead to a non-representative sample of the entire student body, thereby questioning the generalizability of CLA results. However, analyses of sample representativeness show that CLA participants are demographically similar to non-participants (e.g., in average SAT scores, percentage of minority students, and percentage of women) and that controlling for student characteristics has negligible effects on the relative standings of institutions (Klein, Freedman, Shavelson, & Bolus, 2008), suggesting that CLA results are indeed generalizable to an institution.

Critique: The Ceiling Effect
A number of critics, including college presidents from prestigious colleges such as the Universities of Pennsylvania and Virginia argue their students come in as freshmen already at the top of CLA scores. Therefore, there is no point in their institutions using the CLA because their students would show little or no growth on the assessment.

Response: Jeff Steedle, CAE Measurement Scientist in Interview with Neal Conan, NPR, March 22, 2012 on the topic "How Should We Test Students' College Educations?

Steedle: I would want to address Dan's (Dan Barrett, reporter, *The Chronicle of Higher Education*) comments about the possible—what we would call a—ceiling effect, where a Harvard or, we saw recently, a University of Texas might suggest that, well, the reason we don't see large gains is that our students are already coming in toward the top. But indeed in our data, we just don't find evidence consistent with the fact that there's a ceiling effect. We see a normal distribution of the universities, even...among those selective universities, the average gains at those schools are very similar to the average gains at less selective universities. And if it was the case that there was a ceiling effect, we wouldn't see that. We would see smaller average gains at those more selective schools than we do at the less selective schools."

Conan: So what do you conclude from that?

Steedle: We conclude that it's not fair to explain away your scores by claiming that there's a ceiling effect when in fact there's no statistical evidence to support that.

Critique: Students are not motivated.
Low motivation is a persistent threat to the validity of score interpretations, especially for low-stakes tests like the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and the CLA. That is, if examinees are not motivated, their scores will not be accurate reflections of their maximum level of proficiency. To address this concern, studies have been carried out to evaluate the relationship between motivation and performance assessment scores, identify the reasons students are motivated (or not) on performance assessments, and measure differences in motivation on performance assessments observed in low- and high-stakes environments (Steedle, 2010a). Results from these studies show that aggregate student motivation is not a significant predictor of aggregate CLA performance. Therefore, the relative standings of institutions is not affected by student motivation. Moreover, that the types of incentives that students prefer (e.g., money, public recognition) are not related to motivation and performance (Steedle, 2010a).

Critique: As a test of generic skills, CLA results cannot be usefully applied to improve educational programs.

The CLA is a standardized test (i.e., a test administered in the same conditions for all examinees), and it is often the belief that such assessments are not useful for improving classroom instruction. However, there is increasing evidence that performance tasks like those included in the CLA can play an important role in classroom learning and assessment(Chun, 2010).This is important because in order for faculty to take assessment seriously they must view measures as authentic and useful to them in the classroom. Dr. Marc Chun (former Director of CLA Education) has given over 100 faculty academies, including eight in countries besides the United States. Prospects for the development of international performance tasks for AHELO, based on his work to date, appear promising. Appendix B provides some relevant examples of how the CLA has contributed to the improvement of teaching and learning in higher education.

Critique: Generic skills are not independent of discipline-specific knowledge

Critics question whether generic skills like analytic reasoning and problem-solving can be measured independently from discipline-specific contexts. Recent research on this found no significant interaction between CLA Performance Task content and students' fields of study (Steedle & Bradley, 2012). For example, students in the "hard" sciences do no better or worse on a performance task set in the context of a scientific inquiry than they do on a task set in a social science or business context. This finding suggests that generic skills can be measured using complex, authentic assessments without great concern for the potential confounding effect of content knowledge on test performance.

conclusion

Generic skills can be identified and measured. For nearly a decade, the CLA has been measuring higher-order thinking skills that are important to all students regardless of their academic background (Klein, Benjamin, et al., 2007), making it an ideal measure of generic skills for the AHELO project. There is considerable research support for the CLA but more research is needed on a variety of issues. As is the case with other testing organizations, CAE researchers carry out much of the research on reliability and validity issues. However, CAE-based research continues to be published in peer reviewed journals. Moreover, CAE's policy is to provide extensive data sets, when requested, to independent researchers to carry out their own studies (cf. Arum, et al., 2012; Arum & Roksa, 2011). Finally, CAE welcomes the independent studies and reports on the use of CLA results noted in Appendices A and B. Additionally, CAE is planning to use the test results of the AHELO Feasibility Study to conduct analyses of the reliability and validity of measures of generic skills in the international arena.

There is a clear need for a standardized, performance-based assessment that measures higher-order skills across all domains, and the CLA delivers this. If an assessment measured domain-specific knowledge, then the capability to collect information pertaining to students' higher-order skills across multiple domains is lost. The Engineering and Economics Strands of the AHELO project assess only the knowledge, skills, and abilities within these specific domains which leads to this issue of generalizability.

In sum, the Generic Skills Strand is important to AHELO, and performance assessment is well aligned with the educational traditions of most participating AHELO countries that have not experienced the regimen of multiple-choice tests, the testing paradigm of choice by a small number of countries led by the United States. Open-ended essays are the preferred mode of tests in most countries. With technological advances, performance assessments can be scored, analyzed, and reported back to students and their schools in a cost effective and accurate manner. We can move beyond the multiple-choice paradigm which, while still important for a variety of purposes, is not appropriate for benchmarking generic skills and stimulating improvements in the teaching and learning of these skills. Hopefully, as AHELO goes to scale, faculty representatives from participating countries will propose new topics for developing performance assessments in order to move beyond the current set of US-based performance tasks.

References

American Philosophical Association. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction "the delphi report"*. Committee on Pre-College Philosophy. Millbrae, CA: The California Academic Press.

Arum, R., Cho, E., Kim, J., & Roksa, J. (2012). *Documenting uncertain times: Post-graduate transitions of the academically adrift cohort*. Brooklyn, NY: Social Science Research Council.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Association of American Colleges and Universities, & Council for Higher Education Accreditation. (2008). *New leadership for student learning and accountability: A statement of principles, commitments to action*. Washington, DC: Association of American Colleges and Universities and the Council for Higher Education Accreditation.

Astin, A., & Lee, J. (2003). How risky are one-shot cross-sectional assessments of undergraduate students? *Research in Higher Education*, 44, 657-672.

Banta, T. W. (2007, January 26). A warning on measuring learning outcomes. *Inside Higher Ed*.

Banta, T. W. (2008). Editor's notes: Trying to clothe the emperor. *Assessment Update*, 20(2), 3-4, 15-16.

Banta, T. W., & Pike, G. R. (2007). Revisiting the blind alley of value added. *Assessment Update*, 19(1), 1-2, 14-15.

Becker, G. (1964). *Human capital*. Chicago: University of Chicago Press.

Benjamin, R. (2012). *The new limits of education policy*. London: Edward Elgar.

Bok, D. (2005). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton, NJ: Princeton University Press.

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.

Bransford, J., Brown, A., & Cocking, R. (Eds.). (2000). *How people learn*. Washington, DC: The National Academy Press.

Business-Higher Education Forum. (2004). *Public accountability for student learning in higher education: Issues and options*. Washington, DC: American Council on Education.

Campbell, D. T. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education*, 47(1), 1-32.

Chun, M. (2010). Taking teaching to (performance) task: Linking pedagogical and assessment practice. *Change*, 42(2), 22-29.

Council for Aid to Education. (2011). *Module a progress report: Milestone 3*. Paris: Organisation for Economic Co-operation and Development (OECD).

Council for Independent Colleges. (2008). *Evidence of learning: Applying the Collegiate Learning Assessment to Improve teaching and learning in the Liberal Arts Colleges Experience*. Washington D.C.: Council for Independent Colleges

Council for Independent Colleges (2011). *Catalyst For Change: The CIC/CLA Consortium*. Washington D.C.: Council for Independent Colleges.

Dewey, J. (1910). *How we think*. Boston, MA: D.C. Heath.

Douglass, J. A., Thomson, G., & Zhao, C. M. (2012). *Searching for the Holy Grail of Learning Outcomes*, Berkeley, CA: Center for Studies in Higher Education, Research & Occasional Paper Series. CSHE3,12.

Educational Policies Commission. (1961). *The central purpose of American education*. Washington, DC: National Education Association.

Elliot, S. (2011). *Computer-assisted scoring for performance tasks for the CLA and CWRA*. New York: Council for Aid to Education.

Ennis, R. H. (2008). Nationwide testing of critical thinking for higher education: Vigilance required. *Teaching Philosophy*, 31(1), 1-26.
Erisman, W. (2009). *Measuring student learning as an indicator of institutional effectiveness: Practices, challenges, and possibilities*. Austin, TX: Higher Education Policy Institute.
Erwin, T. D. (2000). *The npec sourcebook on assessment, volume 1: Definitions and assessment methods for critical thinking, problem solving, and writing*. Harrisonburg, VA: Center for Assessment and Research Studies, James Madison University.
Erwin, T. D., & Sebrell, K. W. (2003). Assessment of critical thinking: Ets's tasks in critical thinking. *The Journal of General Education*, 52(1), 50-70.
Garcia, P. (2007). *How to assess expected value added: The CLA method*. Paper presented at the California Association of Institutional Research Conference, Monterey, CA.
Gardner, H. (2006). *Multiple intelligences*: New horizons. New York: Basic Books.
Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw Hill.
Hacker, A. (2009, February 24). Can we make america smarter? *New York Review of Books*.
Hardison, C. M., & Vilamovska, A.-M. (2008). *The Collegiate Learning Assessment: Setting standards for performance at a college or university* (No. PM-2487-1-CAE). Santa Monica, CA: RAND.
Hart Research Associates. (2006). *How should colleges prepare students to succeed in today's global economy? - based on surveys among employers and recent college graduates*. Washington, DC: Hart Research Associates.
Hart Research Associates. (2009). *Learning and assessment: Trends in undergraduate education - a survey among members of the association of American colleges and universities*. Washington, DC: Hart Research Associates.
Heckman, J., & Krueger, A. B. (2003). *Inequality in america: What role for human capital policies?* Boston: MIT Press.
Heinrichs, W. L., Lukoff, B., Youngblood, P., Dev, P., Shavelson, R., Hasson, H. M., et al. (2007). Criterion-based training with surgical simulators: Proficienty of experiences surgeons. *Journal of the Society of Laparoendoscopic Surgeons, 11(3)*, 273-302.
Hosch, B. J. (2010). *Time on test, student motivation, and performance on the Collegiate Learning Assessment: Implications for institutional accountability*. Paper presented at the Association for Institutional Research Annual Forum, Chicago, IL.
Hutchings, P. (2010). *Opening doors to faculty involvement in assessment*: University of Illinois at Urbana-Champaign: National Institute for Learning Outcomes Assessment.
Immerwahr, J. (2000). *Great expectations: How the public and parents--white, african American, and hispanic--view higher education*. San Jose, CA: The National Center for Public Policy and Higher Education.
Klein, S. (1996). The costs and benefits of performance testing on the bar examination. *The Bar Examiner*, 65(3), 13-20.
Klein, S. (2002). Direct assessment of cumulative student learning. *Peer Review*, 4(2/3), 26-28.
Klein, S. (2008). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. In D. Nolan & T. Speed (Eds.), *Probability and statistics: Essays in honor of David a. Freedman* (Vol. 2, pp. 76-89). Beachwood, OH: Institute of Mathematical Statistics.
Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review*, 31(5), 415-439.
Klein, S., Freedman, D., Shavelson, R., & Bolus, R. (2008). Assessing school effectiveness. *Evaluation Review*, 32(6), 511-525.
Klein, S., Kuh, G. D., Chun, M., Hamilton, L., & Shavelson, R. (2005). An approach to measuring cognitive outcomes across higher education institutions. *Research in Higher Education*, 46(3), 251-276.
Klein, S., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., et al. (2009). Test validity study (tvs) report. Supported by the fund for the improvement of postsecondary education. from http://www.voluntarysystem.org/docs/reports/TVSReport_Final.pdf

Klein, S., Shavelson, R., & Benjamin, R. (2007, February 8, 2007). Setting the record straight. *Inside Higher Ed.*
Klein, S., Steedle, J., & Kugelmass, H. (2009). *CLA Lumina longitudinal study summary findings*. New York: Council for Aid to Education.
Kuh, G. D. (2006). *Director's message - engaged learning: Fostering success for all students*. Bloomington, IN: National Survey of Student Engagement.
Levy, F., & Murname, R. J. (2004). Education and the changing job market: An education centered on complex thinking and communicating is a graduate's passport to prosperity. *Educational Leadership*, 62(2), 80-83.
Liu, O. L. (2008). *Measuring learning outcomes in higher education using the measure of academic proficiency and progress (mapp)*. (ETS RR-08-47). Princeton, NJ: ETS.
Liu, O. L. (2011a). Measuring value-added in higher education: Conditions and caveats. Results from using the measure of academic proficiency and progress (mapp). *Assessment and Evaluation in Higher Education*, 36(1), 81-94.
Liu, O. L. (2011b). Outcomes assessment in higher education: Challenges and future research in the context of voluntary system of accountability. *Educational Measurement: Issues and Practice*, 30(3), 2-9.
Liu, O. L. (2011c). Value-added assessment in higher education: A comparison of two methods. *Higher Education*, 61(4), 445-461.
McPherson, P., & Shulenburger, D. (2006). *Toward a voluntary system of accountability (vsa) for public universities and colleges*. Washington, DC: National Association of State Universities and Land-Grant Colleges.
Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. San Francisco, CA: Jossey-Bass.
Possin, K. (2008). A field guide to critical-thinking assessment. Teaching Philosophy, 31(3), 201-228.
Report of the cuny task force on system-wide assessment of undergraduate learning gains. (2011).
Shavelson, R. J. (2007a). Assessing student learning responsibly: From history to an audacious proposal. *Change*, 39(1), 26-33.
Shavelson, R. J. (2007b). *A brief history of student learning assessment: How we got where we are and a proposal for where to go nex*t. Washington, DC: Association of American Colleges and Universities.
Shavelson, R. J. (2008). *The Collegiate Learning Assessment*. Paper presented at the Forom for the Future of Higher Education, Cambridge, MA.
Shavelson, R. J. (2010). *Measuring college learning responsibly: Accountability in a new era*. Stanford, CA: Stanford University Press.
Shavelson, R. J. (2011). *An approach to testing and modeling competence*. Paper presented at the Modeling and Measurement Competencies in Higher Education.
Shavelson, R. J., & Huang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change*, 35(1), 10-19.
Shavelson, R. J., Klein, S., & Benjamin, R. (2009, October 16). The limitations of portfolios. *Inside Higher Ed*.
Silva, E. (2008). *Measuring skills for the 21st century*. Washington, DC: Education Sector.
Simon, H. (1996). *The sciences of the artificial*. Boston, MA: MIT Press.
Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, 15, 201-293.
State Higher Education Executive Officers. (2005). *Accountability for better results: A national imperative for higher educatio*n. Boulder, CO: State Higher Education Executive Officers.
Steedle, J. T. (2010a). *Incentives, motivation, and performance on a low-stakes test of college learning*. Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.
Steedle, J. T. (2010b). On the foundations of standardized assessment of college outcomes and estimating value added. In K. Carey & M. Schneider (Eds.), *Accountability in American higher education*. New York, NY: Palgrave Macmillan.
Steedle, J. T. (2011 online first). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education*, 1-16.

Steedle, J. T., & Bradley, M. (2012). *Majors matter: Differential performance on a test of general college outcomes*. Paper presented at the Annual Meeting of the American Educational Research Association, Vancouver, Canada.

Steedle, J. T., Kugelmass, H., & Nemeth, A. (2010). What do they measure? Comparing three learning outcomes assessments. *Change*, 42(4), 33-37.

The New Commission on the Skills of the American Workforce. (2006). *Tough choices or tough times*. Washington, DC: National Center on Education and the Economy.

The Secretary's Commission On Achieving Necessary Skills. (1991). *What work requires of schools: A scans report for america 2000*. Washington, DC: U.S. Department of Labor.

Thomson, G., & Douglas, J. A. (2009). *Decoding learning gains: Measuring outcomes and the pivotal role of the major and student backgrounds*. Berkeley, CA: Center for Studies in Higher Education.

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. Higher education*. Washington, DC.

Zahner, D., Ramsaran, L. M., & Steedle, J. T. (2012). *Comparing alternatives in the prediction of college success*. Paper presented at the Annual Meeting of the American Educational Research Association, Vancouver, Canada.

Appendix A: CLA Critiques and Responses Summary

TOPIC	CRITIQUE	RESPONSE
General/ Background	• The CLA is a one-size-fits all measure designed for accountability only (Douglass, et al., (2012). • The CLA crowds out more nuanced assessments such as portfolios, surveys (Douglass, et al., 2012). • The CLA is designed to create a ranking system (Douglass, et al., 2012).	• Seven Red Herrings on Assessment in Higher Education (Benjamin, 2012) The CLA program rejects one-size-fits all measures. The CLA program is opposed to ranking systems of colleges and universities. Appropriate standardized tests permit inter-institution comparison are necessary but not sufficient. Comparison is needed to frame within-institution formative assessments. Formative assessments are supported, indeed undertaken by the CLA Education program as well. • CLA background and context (Klein, 2002; Shavelson, 2007a, 2007b, 2010; Steedle, 2010b) • CLA constructs (Shavelson & Huang, 2003)
Value-Added	• CLA value-added scores are not reliable (Banta, 2008; Banta & Pike, 2007). • Value-added scores should account for more than just the SAT (e.g., age, race, sex). • Different value-added models can produce very different results (Liu, 2011c). • Value-added approach weakens correlations (Kuh, 2006).	• Average CLA scores are highly reliable, especially when the unit of analysis is the institution (Freshman = .94; Seniors = .86) (Klein, Benjamin, et al., 2007; Klein, et al., 2005). • Adding age, race, and sex to the model does not affect value-added results. Since the variables are correlated with each other, the estimates are less precise due to multicollinearity (Klein, et al., 2008). • It is not true that different value-added models produce different results, as long as you are controlling for EAA (Steedle, 2011 online first).
Reliability	• Tests measure the same thing if they are highly correlated (Belgian NPM and TAG).	• CLA and multiple-choice tests like CAAP are highly correlated, but many tests of obviously different constructs are also highly correlated (e.g., science and reading). Just because the tests are correlated, that does not necessarily mean they are measuring the same thing (Klein, Liu, et al., 2009; Steedle, et al., 2010).
Motivation	• Student motivation affects CLA scores (Banta, 2008; Liu, 2011c, Douglass, et al., 2012).	• Aggregate student motivation is not a significant predictor of aggregate CLA performance. It does not invalidate the comparison of schools based upon CLA scores. The types of incentives that students prefer (e.g., money, public recognition) are not related to motivation and performance (Steedle, 2010a).
Validity	• Tests like the CLA do not measure every important outcome of higher education. "… standardized measures currently address only a small part of what matters in college" (Association of American	• Critical thinking may only be a small part of what students are expected to learn in college. However, it is still a very important skill. In fact, many colleges have a set of general learning outcomes for all students

(Validity - continued)	Colleges and Universities & Council for Higher Education Accreditation, 2008, p. 5). • The CLA tests primarily entering ability (e.g., when the institution is the unit of analysis, the correlation between scores on these tests and entering ACT/SAT scores is quite high, ranging from .7 to .9), therefore differences in test scores reflect individual differences among students taking the test more accurately than they illustrate differences in the quality of education offered at different institutions (Banta, 2007). • CLA tasks are not content neutral, thus they disadvantage students specializing in some disciplines (Banta, 2007, 2008; Banta & Pike, 2007) (Douglass, et al., 2012). • Contain questions and problems that do not match the learning experiences of all students at any given institution (Banta, 2007; Douglass, et al., 2012). • Measures at best 30% of the knowledge and skills that faculty want students to develop in the course of their general education experiences (Banta, 2007). • CLA is not a valid assessment. • The CLA is highly inter correlated with the SAT (Douglass, et al., 2012) and therefore not credible.	regardless of their concentration, and critical thinking and writing frequently occur at the top of the list (Hart Research Associates, 2009). • Although the CLA is correlated with entering academic ability, it does not test the same constructs as college entrance exams like the SAT and ACT (Klein, Shavelson, et al., 2007; Zahner, et al., 2012). • There is no interaction between CLA task content and field of study (Klein, Shavelson, et al., 2007; Steedle & Bradley, 2012). • Isn't it excellent that an assessment measures 30% of the knowledge and skills that faculty want? What assessment out there measures more than this? (Klein, Shavelson, et al., 2007). • The CLA has face validity (Hardison & Vilamovska, 2008, pp. 107-109). • The CLA is sensitive to differences between freshmen and seniors (Klein, Benjamin, et al., 2007). • The most accurate prediction of college senior GPA was achieved using high school GPA plus CLA scores (predictive validity) (Zahner, et al., 2012). • Evidence of CLA reliability, convergent validity, and differences between freshmen and seniors (Klein, Liu, et al., 2009). • Correlations between CLA and the National Survey of Student Engagement (Carini, et al., 2006). • Correlations among Performance Tasks and the GRE (convergent validity) (Klein, et al., 2005).
Test Administration	• Test administration procedures need to be standardized because they appear to influence student motivation and test performance (Hosch, 2010).	• This is a legitimate concern, but we do not have any research published on this issue.
Sampling	• Cannot be given to samples of volunteers if scores are to be generalized to all students and used in making important decisions such as the ranking of institutions on the basis of presumed quality (Banta, 2007). • Longitudinal and cross-sectional data are not comparable (Garcia, 2007). • Freshmen and seniors in a cross-sectional sample are not similar. • No way to determine whether sample is representative (Douglass, et al., 2012). • Small sample required only valid for small liberal arts colleges.	• CLA participants are like non-participants (in terms of SAT scores, ethnicity, and sex) (Klein, et al., 2008). The degree of representativeness is checked with that of the overall student body. • Provides some arguments against longitudinal approach (e.g., expensive, large attrition, and students not progressing in their studies at the same rate within and across schools). May be providing biased results. We can never really know which approach is better or worse. The approaches have different pros and cons and neither is likely to produce an unbiased result (Klein, et al., 2008). • Freshmen and seniors do not differ much from each other except for their CLA scores

(Sampling - continued)		(Klein, et al., 2008). • Cross-sectional provides comparable results to longitudinal (Klein, Steedle, & Kugelmass, 2009). • Small sample is adequate for large universities who, however, may test more students to drill down to departments and programs.
Pedagogy	• Faculty may narrow the curriculum to focus on test content (Banta, 2007) (Douglass, et al., 2012).	• How the CLA relates to what occurs in the classroom and if the CLA results can be used to improve pedagogy (Chun, 2010). • CLA focuses on broad competencies that are mentioned that cut across academic disciplines. Faculty cannot "teach to the test" (Klein, Shavelson, et al., 2007).
Miscellaneous Articles	• Study focused on ETS's Tasks in Critical Thinking and its relation to General Education coursework (Erwin & Sebrell, 2003). • Cross-sectional assessments are difficult to interpret because they inevitably reflect characteristics of the same students when they first entered college; variation is attributable to entering freshman characteristics not institutional policies or practices (Astin & Lee, 2003). • Cannot make America smarter, so there is no need for measures such as the CLA (Hacker, 2009).	• Measuring learning outcomes in higher education (Liu, 2008, 2011a, 2011b). • Limitations of portfolios (Shavelson, Klein, & Benjamin, 2009, October 16) • (Klein, 2002). • Machine-scoring of assessments (Klein, 2008). • Performance testing on the bar exam (Klein, 1996). • Recommends cooperation by critical-thinking faculty and administrators if there is less comparability and deeper transparency of tests (Ennis, 2008). • Non-technical guide to popular methods and tests for assessing how well students acquire critical thinking skills in school and college(Possin, 2008). • Comparison of the methodology and potential uses of three tools for measuring learning outcomes: the CLA, the National Survey of Student Engagement (NSSE), and the University of California's Undergraduate Experience Survey (UCUES) (Thomson & Douglas, 2009). • Examination of the strengths and limitations of some common approaches to measuring student learning outcomes (Erisman, 2009). • Recommendation of the CLA for formative assessment use (Hutchings, 2010). • Comparison of the CLA, CAAP, and Academic Profiles (Report of the cuny task force on system-wide assessment of undergraduate learning gains, 2011). • Use of the CLA as a dependent variable (Arum & Roksa, 2011).

Appendix B: Examples of How the CLA Can Contribute to the Improvement of Teaching and Learning in Higher Education

Of course, the CLA is a testing program. Equally, however, it can be viewed as an instrument for reform of teaching and learning in higher education. It is important to give examples of what this means because the word "assessment" places the CLA in a box occupied by many other assessments, including multiple-choice tests. When examined for its contributions to teaching and learning, the CLA is in a league of its own. Here is a template that indicates how an institution might respond to the initial institutional level CLA scores followed by illustrations of how administrators and faculty are benefiting from using the CLA along with other measures related to student learning. These illustrations are offered as early examples of productive uses of the CLA.

From the Institution to the Classroom: The CLA Comparison Strategy

1. The CLA's single global institutional score is based on the average performance of the sample of freshmen and senior students taking the CLA. An institution's score is presented in comparison to other similarly selective participating institutions. To account for variation in competencies the students bring to college, the CLA institution scores are adjusted for the SAT scores of the participating students. The CLA scores, then, reflect the amount of value-added improvement in performance between the freshman and the senior-year graduating students. When the scores of all institutions taking the CLA are placed in a regression equation, the institutions cluster along a straight line. More specifically, a college can be compared against the performance of other colleges with similar average SAT scores.

   The first time the institution tests, CLA results provide faculty and administrators a benchmark, a signal about where their institution stands. There is up to a 2.0 standard deviation in estimated CLA gains between similarly situated institutions. In other words, there are very large differences in CLA scores between institutions that accept students with similar incoming cognitive ability. This means there is a large canvas for studying best practices in the institutions that perform better than the equation predicts as opposed to those that perform worse. The question then is what should the faculty and administrators of institutions do to improve the degree of their value added? That leads to the following subsequent steps.

2. Correlate inputs, processes, and outputs. A logical next step is for the college's institutional research office to correlate the inputs and processes (or their proxies such as class size, expenditures per pupil, incoming SAT scores of the freshmen, per student endowment expenditures, etc.) with outputs of undergraduate education such as retention and graduation rates and, of course, CLA outcomes and other measures of learning. The goal here is to develop an efficient description of the factors that correlate with CLA results.

3. Conduct in-depth analysis. While the institutional score signals the place of the college compared to other colleges administering the CLA, college administrators and faculty members will want to know more about the relative contributions to that score by colleges (if the institution is a university) or by certain departments or programs (if the institution is a college).Which departments or programs, for example, are particularly strong or weak contributors to their CLA results?

4. Conduct audit of existing assessments. There is a saying in the assessment world that a curriculum is determined by what faculty test for. Thus it will be useful to understand the extent to which faculty are using multiple-choice or essay tests in their classrooms. Are the tests given measuring

what is important such as critical thinking, problem solving or analytical reasoning? How well are the students doing on current tests?

5. Examine best practices found to produce good CLA results. Many colleges participating in the CLA are working together in consortia of similar institutions. They are highlighting and sharing best practices that are correlated with noteworthy CLA scores. For example, it appears that schools that require more analytic-based writing do better on the CLA than those that do not.

6. The most important step: get published CLA Performance Tasks into the hands of the faculty so that they can:
    a. Use them in their classroom where they have greater knowledge of the strengths and weaknesses of their students;
    b. Develop Performance Tasks that are based on the scoring guide of the published tasks;
    c. Choose case studies and problems for text material that is congruent with the documents in the CLA Performance Tasks rather than the content dominated textbooks extant;
    d. Adopt a student-centered approach to teaching that calls for much more analytic-based writing on the part of the students and diagnostic feedback to the student about how they can improve their performance.

In sum, the above steps comprise an early version of what we hope will become a reinforcing system of continuous improvement of teaching and learning.[1] The institution's global score provides a critical signal that triggers an internal focus on what correlates with the score. It does not really matter where the institution is on the initial test administrations. The important questions become related to (a) understanding what led to those results and (b) deciding what improvement goals might make sense for the future.

Below are a few links illustrating how the CLA has contributed to the Improvement of Teaching and Learning in Education.

- http://www.uky.edu/IRPE/assessment/Sizzle_June_2009.pdf
- http://21k12blog.net/2010/10/19/performance-task-assessment-and-teaching-learning-from-chun-and-clacwra/
- http://21k12blog.net/2010/02/24/excellent-cwra-info-session-at-nais-true-21st-c-assmt-naisac10/
- http://www.teaglefoundation.org/liblog/entry.aspx?id=252
- http://teachingatfsu.com/?p=40

In addition see three reports summarizing work of faculty in a consortium of 47 private liberal arts colleges led by the Council for Independent colleges to share best practices that improve student learning growth; including:

- Evidence of Learning: Applying the Collegiate Learning Assessment to Improve Teaching and Learning in the Liberal Arts College Experience.
- Catalyst for Change: The CIC/CLA Consortium.
- An Analysis of Learning Outcomes of Underrepresented Students at Urban Institutions.

[1] This is precisely what higher education has in the research realm. Through peer review research has a public face that encourages and requires researchers to respond to criticism and evaluate the claims of other researchers: in short, engage in a never ending process of continuous improvement. If we followed the above steps for undergraduate assessment, we could hope to eventually also create a continuous system of improvement of teaching and learning.

The New Limits of Education Policy

Roger Benjamin

Introduction

The main focus of both political parties is wrong. It's not about jobs. It is about education. In today's knowledge economy the country with the highest quality postsecondary education will generate the new ideas that have economic value, the entrepreneurs who see around the corners, and the new markets for new products. High school dropouts do not have the skills to participate in the knowledge economy. Yes, many high school dropouts are employed, but in minimum wage jobs rather than high value-added economic activities. This is the unacceptable state our economy and society are in today. Skill deficits are the real problem underlying the high unemployment figure in the United States. When listening to the speeches of public and private leaders one does not get the sense that education policy is central to their agenda. This is a huge concern. The remedy is to insist upon framing education policy along with the institutions and practices it guides in a new way suggesting new paths for action.

The Rationale for Assessment

The assumption that human capital is the only real national or individual resource is becoming ascendant in the United States. That means that education or training is the formal (and really only practical) way for individuals to improve their human capital. But the reality is that we have a very large group of unskilled citizens and high school dropouts for whom we do not offer postsecondary training or education opportunities. U.S. society as a whole cannot be successful in today's knowledge economy with such a large proportion of our population without the skills to contribute to the economy. It should therefore be the right of all citizens to purse some form of postsecondary education or training.

Despite this clear need, there appears to be significant underspending on postsecondary education in the US. The consequence of inadequate revenue is a hollowing out of the sector's infrastructure, particularly those devoted to instruction. Institutions are forced to reduce the number of instructors in the classroom and to eat their capital stock through deferring maintenance and not renewing technology. The inevitable results are a reduced ability of colleges and universities to admit students and a decline in the quality of education they are able to provide. If this situation continues, we risk tipping over into a permanent social and economic crisis.

That is why I argue we need a fundamental debate about higher education in the United States. Why are we underinvesting in higher education when its role in securing our economic and social well being is so evident?

This debate must start by thinking carefully about the balance between the public and private responsibilities for postsecondary education. One may think of goods as a continuum from pure public to pure private goods. In reality, there are few if any pure public goods that warrant that description. Nor are there many pure private goods. Most goods produce some distortion effects or exhibit free rider problems. That has led to the introduction of terms such as quasi-public, quasi-private, or collective goods as descriptors of goods.

Unlike elementary education, which most countries would accept has large positive social benefits and therefore merits extensive public support, the situation in postsecondary education is less clear. There are significant financial and non-financial returns to the individual, which is the usual justification for students paying at least part of the cost of their higher education. But it would not be accurate to label postsecondary education a pure private good. There are clear public good dimensions to research and teach toward in postsecondary education, justifying the infusion of public funds into the sector. This social dimension to

higher education was recognized in the United States when the land grant university was created to permit universal access to all citizens who meet admissions requirements through the Morrill Act by Congress in 1866.

I suggest viewing postsecondary education's problems through the common pool concept because the reality is that there is not sufficient public or private spending on the postsecondary education good. CPPs develop whenever a group depends on a public good that everyone uses but no one owns, and where one person's use affects another person's ability to use the good. The result is that either the population fails to provide the resources, over consumes these resources, and/or fails to replenish them. Another important characteristic includes confusion about property rights. In this case, it is important to discern who is responsible for seeing to the underinvestment in higher education.

There also is widespread confusion about what we should do about the lack of investment in higher education. In the United States there is a historic legacy that treated postsecondary education through the Land Grant model in recognition of its large social benefits. In recent decades, accelerated by the effects of the recent recession, net tuition revenues outstrip local and state subsidies in over one half of the states.

We need a public debate about the issues the postsecondary sector confronts, including how we treat the sector conceptually.

A variety of indicators may be presented to make the case for alarm. First, there are approximately 47 million high school dropouts, about one-sixth of the total U.S. population. These citizens do not have sufficient qualifications to enter postsecondary education. Moreover, within these dropout statistics rest America's ethnic and racial divisions. A disproportionate number of African Americans and Latinos make up the high school dropout category while the Latinos, alone, constitute approximately 80 percent of the potential growth in postsecondary education over the next several decades. Because the college-going rate of Latino and African Americans is less than half of non-Hispanic whites and Asian Americans, there is a serious shortfall in postsecondary education attendance. Of course, political leaders, private sector leaders, education officials throughout the K-16 system, and many interest groups and concerned individuals often point to the seriousness of the problem as signified by the size of the high school dropout number. Yet, the number of high school dropouts continues to rise. Why is this issue not a critical public policy issue of the highest importance?

The postsecondary access deficit problem is exacerbated by poor preparation and graduation rates of those who go to college. Forty percent of students who enter college require remediation because they do not have minimum college readiness levels of proficiency in reading, writing, and mathematics. Despite the doubling of spending on K-12 education since the mid 1970s the scores of nationally normed tests such as the National Literacy test, the National Assessment of Educational Progress (NAEP), and the SAT and ACT have all been flat or declining for the past 25 years. Eighty percent of high school students recognize that there are gaps in their preparations for college and 65 percent of college instructors report that high schools are not adequately preparing their graduates for college course work. Employer surveys report widespread dissatisfaction with the critical thinking, analytical reasoning, problem solving, and writing skills of the potential employees they screen. Funding has shifted from public to private sources over the past three decades. Typically, states are mandated to fund K-12 education, prisons, and health care. The result is higher-education budget allocations are a source of funding these other programs. Finally, there are serious questions about the learning gains of students who complete college at all levels indicating that the quality of the learning produced at the nation's campuses is less than we would have hoped.

The New Place of Education Policy

If human capital is the principal national resource and education the formal venue to preserve and develop it, education and education policy and research should be recognized as critical to the success of the other main policy domains of concern to the State such as health, the economy, energy, and national security. The problem is there is a disconnect between what should be the case and current reality. The policy priorities of the State have not caught up with this new reality. In the United States, and many countries, education is not a high priority at the national level. For example, the budget for postsecondary education is about $132 million out of a federal research budget

for non-defense research of $150 billion, a rounding error. If one examines the top 100 social science journals, there are none on education research. Education policy as a field of study is not present in main stream economic or political science departments in universities. Nor is it seen as a key subject in public policy schools. Education research is a marginal subject in colleges of education which are themselves marginalized in research universities. Finally, GRE scores of entering graduate schools in education are at the bottom of 28 major academic disciplines.

Coming Changes

Because consensus about the centrality of the human capital concept is likely to be soon shared by most private and public leaders they will also see the disconnect between the marginal attention given to education policy and research, and education institutions compared with defense, health and other policy areas and what should be the case. Education policy should be a much higher national priority in order to develop out of the box solutions to improve the K-16 education system, the formal venue to preserve and enhance human capital.

Additionally, rising costs, declining resources and the emergence of Internet-based education solutions add up to a set of disruptive forces that will require significant redesign of traditional postsecondary institutions. This leads to an inevitable growing role for assessment in postsecondary education, a controversial prediction. Why will assessment of student learning increase?

- The assumption that students arrive ready for college is no longer warranted which means benchmarking student learning levels of incoming students at the point of college admissions is needed.
- Evidence suggests that we can not assume the results of a four-year education are acceptable. Low retention and graduation rates coupled with surveys and recent studies of the quality of student learning also suggests benchmarking the quality of student learning achieved by graduating students is needed.
- To achieve greater efficiencies and effectiveness, for-profit colleges have moved to a competency-based model of learning that requires outcomes assessment. There is a strong argument for using independent, third party-based standardized assessments of the efficacy of competency-based based education outcomes.
- Uncertainty about the quality of online courses in "traditional" colleges places a greater focus on assessing the student learning results as well.
- Rising costs and declining resources make restructuring of academic programs inevitable because that is where the bulk of resources any college has resides. In the absence of a metric evaluating the impact of budget cuts in selected departments or programs, it is not possible to gauge their effects on student learning. Student learning outcomes are the most logical metric to use.
- Just placing a spotlight on student learning outcomes with assessment-based evidence will generate more scrutiny of the cost and quality of undergraduate education. This is important for institutions of higher education that have few market or market-like mechanisms to discipline costs.
- Most importantly, student learning outcomes need to be measured to provide faculty with the evidence they need to assist them in improving teaching and learning. This is a goal widely shared in the academy.

Here, then, are three factors that make greater transparency of undergraduate education's learning outcomes inevitable.

First, the increasing recognition of the critical importance of human capital will translate into increased external demands for oversight of the learning results educational institutions produce.

Secondly, the cost problem will combine with declining resources to compel re-allocation of resources from low to high priorities. Escalating student loan debt, now close to one trillion dollars in aggregate in the United States, is a product of the cost and declining revenue trends. The negative social-economic implications of this level of student loan debt has emerged as an important issue in its own right and appear to be a galvanizing force destined to place potent pressure on postsecondary institutions and systems to cut costs and for local, state, and federal governments to provide greater aid for students.

Thirdly, the drive for equality for all ethnic/racial groups in a liberal democracy like the United States will accelerate the need for objective information useful to close the gaps in student learning results between ethnic and racial groups.

Impetus for Assessment

The initial impetus for assessment will come from external forces. This is because institutional arrangements in colleges and universities are highly institutionalized and therefore unlikely to be challenged initially from within the academy. Postsecondary institutions have substantial autonomy from the governing bodies they report to. Decentralized, department-based governance delegates substantial decision-making authority to faculty regarding what subjects should be taught and whether and how students should be assessed. Therefore, assessments need to be developed that are deemed to be authentic by the faculty and of direct use to them in the classroom.

One task to undertake is to eliminate the red herrings about assessment in higher education. There are shibboleths that are regarded as self evident truths about assessment in higher education. In fact, they are convenient excuses to not assess student learning outcomes (see appendix for list).

Conclusion: What Should Happen

In the absence of effective institutional redesign strategies made credible with evidence, the best prediction is that the next decade will be one of turmoil with continued rising costs, access deficits, and a stagnating "traditional" postsecondary sector. External demands for assessment will increase but faculty will resist the calls, in particular for more external-based accountability. The decade will feature disarray with few winners, a few selective colleges and research universities. The common pool problem would then become a permanent crisis.

Just as with other major issues such as health care or national defense during the Cold War there should be a major national debate about the common pool problem recognizing the importance of the human capital argument and the urgency it poses to the welfare of the American State. The result should be a strategic repositioning of the mission of the university and a changed incentive system of faculty to encourage improvement of teaching and learning. Greater evidence-based decision making, including assessment of student learning, should be encouraged in order to calibrate the repositioning of the postsecondary sector that will occur over the next two decades.

The American university should restructure its mission to focus on improving the K-16 education system, the main lever for human capital development soon to become the principal priority of the State to preserve and enhance. Higher education must set the performance objectives for students to achieve in order to received a B.A. degree. This is necessary information for parents, teachers, and students in the K-12 space; they need to know what basic goals they should set for students to achieve over the course of their educational career. Otherwise, how will they know what skills to develop? Higher education leaders must also exhibit stronger leadership in setting curricula and pedagogical expectations and standards in teacher education. This will significantly improve K-12 education. Finally, assessments improved and made more accessible through technological advances, will become an essential formative and summative feedback tool necessary for raising standards and setting criteria for the entire K-16 education system.

Appendix

- Since we cannot measure all of education, it is impossible to measure any part of it. *This logical fallacy ignores the fact that it is possible to measure key elements of education.*

- Comparison between institutions is not possible or, in any case not warranted. *Comparisons between similar institutions show up to 2.0 standard deviation differences in value-added improvement to their student learning outcomes. This demonstrates there are best practices to evaluate. And the between institution comparisons permit faculty and administrators to place their institution's student learning results in a larger context.*

- All standardized tests are bad or not needed.
  All standardized tests are not alike. Performance assessments are different that multiple-choice tests. They are tests faculty find authentic. Appropriate standardized tests are necessary to place within institution assessment results in context. Without this ability individual institutions will remain isolated silos, one from the other, without any ability to benchmark how well they are doing.

- One-size-fits-all measures to compare institutions are inappropriate.
  Yes. Colleges and universities are far too complex to be accurately characterized by a single test score. Standardized tests must be combined with suitable formative, within institution tests to provide a full picture of the progress of student learning success in a college.

- Content is what is important in undergraduate education.
  Of course content will always remain important. However, in the knowledge economy it is crucial that students be able to access, structure and use information. When students can Google for facts they need to be able to apply what they know to new situations.

- If colleges engage in systematic assessment of student learning outcomes, the results will be used by authorities to control and punish institutions. *We should use the same concept of peer review used to govern the progress of research for use of student learning outcomes data. Results of testing should remain under the control of the testing institution's leadership. Carefully worked out protocols for public reporting should be developed by colleges and the authorities they report to.*