

**On the Foundations of Standardized Assessment of College Outcomes and Estimating Value Added**

Jeffrey Steedle  
Council for Aid to Education

[jsteedle@cae.org](mailto:jsteedle@cae.org)

**Draft: Please do not cite without permission from the author**

Prepared for the American Enterprise Institute Conference, "Increasing Accountability in American Higher Education"  
November 17, 2009

The collected papers for this conference can be found at [www.aei.org/event/100134](http://www.aei.org/event/100134).

**Draft: Please do not cite without permission from the author**

The notion of holding institutions of higher education accountable for student learning resonates with many during these times of financial constraint and increasing global competition. As the commission appointed by former Secretary of Education Margaret Spellings stated, “...parents and students have no solid evidence, comparable across institutions, of how much students learn in colleges or whether they learn more at one college than another. Similarly, policymakers need more comprehensive data to help them decide whether the national investment in higher education is paying off and how taxpayer dollars could be used more effectively.”<sup>1</sup> Furthermore, business leaders are calling for greater accountability in higher education to ensure that college graduates have the skills necessary to make contributions in the workplace. In part, these calls are motivated by the realization that economic growth in the United States will likely decelerate because of a major shortfall of skilled workers in the next decade.<sup>2</sup>

In recognition of their responsibility to provide evidence of student learning, many colleges and universities have ramped up their institutional assessment programs and joined in collective efforts such as to the Voluntary System of Accountability (VSA) and the University and College Accountability Network (U-CAN) to share comparative data with parents and prospective students. As explained by the Association of American Colleges and Universities (AAC&U) and the Council for Higher Education Accreditation (CHEA), “...we in higher education must constantly monitor the quality of student learning and development, and use the results both to improve achievement and to demonstrate the value of our work to the public.”<sup>3</sup>

To achieve these goals, institutional assessment programs commonly employ student and alumni surveys, locally developed tests, and portfolios of student work. In recent years, the number of schools also administering standardized tests of general college outcomes has

**Draft: Please do not cite without permission from the author**

increased substantially. Indeed, rigorous efforts to improve student learning require direct and comparative measures of current student achievement and a process for gauging improvement. Standardized tests fulfill these requirements by providing indicators of achievement on important college outcomes for groups of students, signaling curricular strengths and weaknesses, and, by virtue of standardization, providing benchmarks for performance and facilitating institutional comparisons based on contributions to learning. As the National Research Council Board on Testing and Assessment affirmed, “In many situations, standardized tests provide the most objective way to compare the performance of a large group of examinees across places and times.”<sup>4</sup>

Current demands for direct measures of college outcomes are fueled by recognition that commonly used indicators of educational quality (e.g., persistence, degree attainment, *U.S. News and World Report* rankings) are correlated with educational inputs such as students’ entering academic ability levels and institutional resources.<sup>5</sup> Directly measuring student learning (i.e., the “output” of college education) seems to be a natural remedy to this problem, but scores from standardized tests of college outcomes also tend to be highly correlated with entering academic ability. This is especially true when the school (rather than the student) is the unit of analysis, as is the case for many standardized testing programs in higher education.<sup>6</sup> Thus, scores on college outcomes tests may only provide yet another reflection of educational inputs, which tend to paint highly selective institutions in a positive light and discount the educational value of attending less selective institutions.

This difficulty manifests the general problem that student achievement is strongly related to prior achievement and a host of other factors like parental education and income levels.<sup>7</sup> This complicates comparisons between institutions based on student achievement because observed

**Draft: Please do not cite without permission from the author**

differences can be explained away by factors unrelated to institutional quality. To address this difficulty, standardized testing programs in higher education have adopted *value-added modeling*, which affords recognition for schools whose students demonstrate significant learning gains even if those students have not yet reached a desired level of proficiency.<sup>8</sup> Specifically, institutional value-added scores indicate whether the learning gain (or average senior test score, depending on the statistical model) observed at a particular school is commensurate with those at institutions with similar educational inputs. This allows all schools, even non-selective ones, to demonstrate the educational value of attending. For example, a value-added score may indicate that the average difference between freshmen and seniors at a school is substantially greater (or less) than one would typically observe at schools with students of similar entering academic ability.

Value-added modeling has been heralded as an important advancement in educational measurement,<sup>9</sup> and it has unquestionable intuitive appeal, but many have expressed doubts about the statistical dependability of value-added scores. Mainly, critics assert that value-added scores cannot be estimated with sufficient precision to warrant valid interpretations as indicators of relative learning gains. At a more fundamental level, critics question the use of standardized tests of general college outcomes, which are necessary for estimating institutional value-added scores. Their primary contention is that tests of general skills cannot possibly capture the outcomes of higher education in a meaningful way because those outcomes are so numerous, complex, and inextricably tied to a particular field of study or curriculum.

This chapter supplies evidence supporting the statistical precision of institutional value-added scores derived from standardized test results. As it turns out, many statistical criticisms of institutional value-added modeling are based on evidence derived from value-added models that

**Draft: Please do not cite without permission from the author**

treat the student as the unit of analysis rather than the institution. Analyses of data from recent administrations of the Collegiate Learning Assessment (CLA) demonstrate that institutional value-added scores are substantially less prone to measurement error than critics claim, but there is undoubtedly room for improvement. Additional analyses show that the dependability and interpretability of value-added scores are greatly improved using a new statistical model. This chapter concludes with recommendations for administering standardized tests and integrating them into comprehensive systems of accountability and improvement.

First, it is necessary to address the more fundamental concerns about the use of standardized tests of general college outcomes. This chapter provides some history of standardized testing in higher education, summarizes critiques of current standardized testing efforts, and presents arguments for the utility of standardized testing as a component of institutional assessment programs. It is argued here that postsecondary schools should be measuring higher-order skills like critical thinking and writing because teaching those skills is central to their missions, standardized tests provide the capacity to gauge strengths and weaknesses on these skills, and standardized test results (and the tests themselves) have the potential to stimulate improvements in teaching and learning.

## **Standardized Testing in Higher Education**

### *Measuring General College Outcomes*

The contemporary history of measuring general college outcomes using standardized tests reaches back to the first half of the twentieth century.<sup>10</sup> Early highlights from this history include the Pennsylvania Study, which tracked thousands of students from their sophomore year in 1930 to their senior year in 1932 at 45 different colleges using an 8-hour battery of objective

**Draft: Please do not cite without permission from the author**

tests (i.e., multiple-choice, matching, and true/false) measuring “memory, judgment, and reasoning ability through simple recognition.”<sup>11</sup> Students were tested on their knowledge of the natural sciences, social sciences, two foreign languages, the history of ancient and modern civilization, math, and English (the original battery administered to seniors in 1928 was 12 hours long and did not include math or English). The Pennsylvania Study provided evidence of large differences in achievement between the 45 participating institutions and also that number of credit hours in the natural sciences as well as language, literature, and fine arts was strongly related to achievement in those domains, respectively (this was not true of social studies).

In the years to follow, tests were designed to assess greater depth of understanding at schools like the University of Chicago, where multiple-choice and essay tests were utilized to measure students’ abilities to apply knowledge, predict outcomes, and make decisions in unfamiliar situations.<sup>12</sup> Other early standardized testing efforts like the Cooperative Study of General Education, which involved a consortium of schools dedicated to improving general education, were more holistic and included non-cognitive outcomes such as life goals, social understanding, and health.<sup>13</sup>

In the 1930s, the Graduate Record Exam (GRE) gained popularity as an objective test of verbal reasoning and of general education content knowledge in several disciplines (mathematics, physical science, social studies, literature and fine arts, and one foreign language) for students applying to graduate studies, and the GRE Advanced Tests were developed to assess high-level content in specific majors. Though, in the decades to follow, general reasoning supplanted content knowledge as the primary construct measured by the GRE.<sup>14</sup> This move was designed to improve comparability across students and schools by reducing the influence of a school’s curriculum or a student’s chosen coursework on test scores. Along these lines, the GRE

**Draft: Please do not cite without permission from the author**

Aptitude test emerged in 1949 with the now-familiar verbal and quantitative reasoning sections, and GRE Area Tests were first administered in 1954 to measure students' abilities to reason with new information and draw valid inferences in the natural sciences, social sciences, and humanities.

Despite the pervasive preference for inexpensive, easy to administer, objective tests like the GRE Aptitude Test, several standardized testing programs that surfaced in the 1970s resisted reliance on multiple-choice tests.<sup>15</sup> These programs included open-ended tasks requiring students to demonstrate critical thinking, problem solving, and communication skills in real-world scenarios. Early versions of ACT's College Outcomes Measures Project (COMP) exemplified these efforts. COMP was originally a 6-hour test including oral, written, and multiple-choice items related to 15 simulated activities that incorporated multimedia stimulus materials like audio recordings, photographs, and newspaper articles. For instance, one COMP task asked students to choose between two paintings that might adorn a new public library in a rural town and to support that choice. Students received scores in three content domains (Functioning with Social Institutions, Using Science and Technology, Using the Arts) and three process domains (Communicating, Solving Problems, and Clarifying Values). Test results revealed that COMP scores tended to increase from freshman to senior year by an effect size of roughly 0.80 and that COMP could discriminate between college seniors and matched groups of students attending vocational or technical schools.<sup>16</sup> Not surprisingly, given the undesirability of long, complicated, expensive administration and scoring procedures, the test was later shortened to 4.5 hours, and ACT eventually developed a 2.5-hour, multiple-choice version.

In the late 1970s and early 1980s, the notion of using standardized tests for institutional accountability arose in response to the widely held belief that education was the key to

**Draft: Please do not cite without permission from the author**

maintaining a global competitive edge and to growing recognition that improving education required measuring learning.<sup>17</sup> States mandated accountability testing, and some schools responded by administering off-the-shelf tests that were not designed for institutional assessment (e.g., the GRE and the ACT). At the time, COMP was the only test specifically designed to measure general college outcomes and to facilitate group-level comparisons (using a reference sample of 15 representative colleges and universities). Schools used this reference sample as a consistent point for comparison when evaluating student achievement. For example, Ohio University saw its average senior COMP performance increase from the 50th percentile to the 66th percentile during the years following its implementation of new general education requirements.<sup>18</sup>

Institutional assessment efforts were largely supplanted by accreditation during the recession of the early 1990s,<sup>19</sup> but demands for accountability testing in higher education persisted. Recent calls for such testing originate chiefly from leaders in higher education, business, and government.<sup>20</sup> As in the past, most are motivated to ensure that college graduates in the U.S. have the skills necessary to maintain a competitive edge in the global marketplace, but contemporary calls for measuring student outcomes are increasingly motivated by a desire to remedy social inequities. The idea is that accountability testing can inform curricular and pedagogical improvements that will help address the needs of college students coming from backgrounds lacking in educational opportunity, thereby narrowing historical gaps in achievement and graduation rates.

*Contemporary Standardized Tests*

Several standardized tests are currently employed to measure general outcomes in higher education. Prominent examples include the College Basic Academic Subjects Examination

**Draft: Please do not cite without permission from the author**

(CBASE) from the University of Missouri Assessment Resource Center, the Collegiate Assessment of Academic Proficiency (CAAP) from ACT (the successor to COMP), the Measure of Academic Proficiency and Progress (MAPP) from ETS, and the CLA from the Council for Aid to Education (where the author is currently employed). CBASE, CAAP, and MAPP are multiple-choice tests, each with an optional essay. CBASE is a 180-item test that provides scores in English, mathematics, science, and social studies. CAAP is administered in 40-minute modules that assess students in reading, writing, essay writing, mathematics, science, and critical thinking. MAPP is a 2-hour test of reading, critical thinking, writing, and mathematics (a 40-minute abbreviated version is also available). Aggregated student scores from any of these tests may be interpreted as indicators of institutional performance. In addition, these tests provide student-level scores that may be employed for identifying an individual's academic strengths and weaknesses or for selection purposes (e.g., CBASE is used as a qualifying test for teacher education programs).

Students participating in the CLA either take an *Analytic Writing* test, which involves developing and critiquing arguments, or a *Performance Task*, which presents students with a real-world problem, provides them with a library of relevant information (some credible, some unreliable), and asks them to propose and justify a course of action and suggest additional research to address unanswered questions. For example, one CLA Performance Task places students in the role of an advisor to the mayor of “Jefferson.” Students are asked to evaluate claims about proposals for reducing crime (increase the number of police and funding a new drug treatment program). The task provides documents such as memoranda, newspaper articles, crime statistics, and abstracts from research journals. The documents include a mix of credible

**Draft: Please do not cite without permission from the author**

and unreliable evidence that must be evaluated and synthesized in order to draw valid conclusions and make informed recommendations to the mayor.

The CAAP and MAPP critical thinking tests bear some resemblance to CLA tasks (especially the *Critique-an-Argument* task) except that CLA tasks employ an open-ended response format. CAAP and MAPP critical thinking tests consist of several passages, each with associated multiple-choice questions. The passages commonly present an argument (or opposing viewpoints) like one CAAP item that provides the transcript of a debate between a proponent and an opponent of allowing pharmacists to distribute certain prescription medications without a prescription written by a doctor. The associated items ask students to identify assumptions and implications and to evaluate the strength of evidence and claims.

Unlike these other tests, the CLA was designed exclusively for use in institutional assessment programs. CLA tasks are randomly distributed among participating students at each school, and students are allotted approximately 90 minutes to complete their essay responses. CLA scores are said to reflect the integration of critical thinking, analytic reasoning, problem solving, and written communication skills. Students receive their CLA scores, but these scores cannot be used to compare students or to draw conclusions about their abilities. The reason is that, for individual students, CLA scores (and test scores reflecting performance on a small number of open-ended tasks in general) are typically more prone to measurement error than scores from well designed multiple-choice tests incorporating a large number of items. Much of this unreliability is related to the interaction between students and tasks (i.e., students might have gotten different scores had they been assigned different CLA prompts). These measurement errors are substantially reduced at the aggregate level, making it sensible to interpret estimates of institutional performance on the CLA.<sup>21</sup>

*Reporting Test Results*

In a recent survey of AAC&U member institutions, roughly one quarter of schools reported that they administer “standardized national tests of general skills, such as critical thinking.”<sup>22</sup> As noted previously, results from such tests are employed for two purposes: demonstrating learning gains and improving teaching and learning. Some schools only review test results internally, but other schools share test results and instructional best practices as part of a consortium like the 33 member institutions of the Council for Independent Colleges (CIC) that have administered the CLA since 2005.<sup>23</sup> At their most recent meeting, presenters shared curriculum reforms that were influenced by standardized test results, methods of engaging faculty in institutional assessment programs, classroom efforts to assess critical thinking and writing skills, and techniques for analyzing test data.<sup>24</sup> As part of a new Teagle Foundation grant, 47 schools will participate in the CIC/CLA Consortium between 2008 and 2011.

Some schools, like those in the University of Texas system, are required by state accountability systems to report results publicly, but some do so voluntarily as part of efforts to demonstrate their commitments to assessment and improvement. The VSA, developed jointly by the Association of Public and Land-grant Universities (formerly NASULGC) and the American Association of State Colleges and Universities (AASCU), exemplifies voluntary collective efforts toward these ends.<sup>25</sup> As part of the VSA, nearly 300 schools publish a College Portrait based on a common template for sharing institutional information with prospective students and their parents. It includes a Student Learning Outcomes section that provides institutional scores on CAAP, MAPP, or the CLA (some schools have yet to select an assessment or post results). The analogous U-CAN Profiles include a template for reporting standardized test results, but its use is strictly voluntary.

**Draft: Please do not cite without permission from the author**

Christine Keller, Executive Director of the VSA, and John Hammang, Director of Special Projects at AASCU, reported that a “volatile mixture of technical matters, philosophical differences, and political fears” made the decision to include standardized test results in the College Portrait highly contentious.<sup>26</sup> Similar debates continue among testing experts and among faculty at institutions struggling to satisfy demands for greater accountability. The following sections explore the principal critiques of standardized testing in higher education and make the case that, despite their limitations, standardized tests can play an important role in comprehensive institutional assessment programs.

*Common Objections*

Despite the near ubiquitous use of standardized tests in high-stakes admission decisions, standardized tests of general college outcomes have yet to win widespread acceptance and adoption even when they are used for low-stakes purposes like identifying academic strengths and weaknesses of students. Above all, critics argue that results from these tests cannot be interpreted as meaningful indicators of student learning. One aspect of this critique focuses on the context-specific nature of college learning. That is, college learning is so deep and so particular to a field of study that it would be grossly inadequate to measure that learning with a test so general that it can be administered to any student.

For instance, a test of college outcomes for history majors might involve critically analyzing historical documents, drawing conclusions about historical events, and communicating ideas effectively. If history majors are assessed on critical thinking and written communication skills apart from the context of historical analyses, one may not observe the true depth of their learning. Furthermore, a history major’s critical thinking and written communication skills may not generalize to contexts other than historical analyses. Either way, results from standardized

**Draft: Please do not cite without permission from the author**

tests of general outcomes would not adequately reflect the student's learning in the domain that matters most: his or her chosen field of study. As an alternative, critics Trudy Banta and Gary Pike, both faculty members at Indiana University Purdue University Indianapolis, "support a focus on major field assessment, with an emphasis on using student electronic portfolios as the most authentic instrument for demonstrating growth over time."<sup>27</sup>

The other aspect of this critique focuses on the idea that there are many important outcomes of postsecondary education that are not measured by currently-available standardized tests (e.g., ethical reasoning, civic engagement, and intercultural skills). Thus, test scores reflecting only a narrow range of outcomes cannot adequately convey the overall value of attending a particular institution. In this manner, the AAC&U and the CHEA, in a joint statement, affirmed the importance of setting educational goals, gathering evidence of student learning, and demonstrating educational value, but they stopped short of recommending that schools publicly share standardized test results because "standardized measures currently address only a small part of what matters in college."<sup>28</sup> The notion is that it would be unfair to evaluate and compare schools using standardized test scores because unfavorable results could conceal exemplary gains in unmeasured domains.

Critics also take issue with the utility of standardized test results. Aggregate scores, which are intended to support inferences about entire institutions(or programs within institutions), typically serve only as indicators that schools are doing something right or something wrong. Unfortunately, it is not apparent what that "something" is, especially when the test is not directly tied to the curriculum. In other words, standardized tests cannot diagnose the cause of specific academic strengths and weaknesses and therefore cannot prescribe ways to enhance the learning environment at a school. For instance, standardized tests results cannot

**Draft: Please do not cite without permission from the author**

indicate that certain classes need to be added to a school's general education requirements or that instructors need to improve some specific aspect of their pedagogy or that new extracurricular programs may enhance learning.

Other negative reactions against the use of standardized tests arise from faculty who, based on the principle of academic freedom, would resist any program that “interferes” with academic matters, even if the purpose is to improve learning. As former Harvard University president Derek Bok explained, “The principle of academic freedom, originally conceived to safeguard a scholar's right to express unpopular ideas, has been stretched to protect individual professors from having to join any collaborative effort to improve the quality of teaching and learning.”<sup>29</sup> Indeed, some faculty react negatively to accountability and assessment programs because of their associations with the federally-mandated (and often maligned) K-12 accountability system created by the No Child Left Behind Act of 2001 (NCLB).

In sum, critics point out that standardized tests do not provide a complete solution to the challenges of institutional assessment. Tests of general college outcomes do not measure depth of learning in students' respective fields of study, and they do not measure every important outcome of undergraduate education. Moreover, results from standardized tests do not diagnose the specific causes for academic strengths and weaknesses, and the introduction of standardized tests, despite claims to the contrary, could have undesirable impacts on the curriculum. Notwithstanding these limitations and concerns, standardized tests may still fulfill an important role in institutional assessment and also have the potential to effect positive changes on teaching and learning. The following section explores these possibilities by examining reasons that postsecondary schools choose to administer standardized tests of general college outcomes.

*Responding to Objections*

Testing proponents (and testing agencies) should be quick to agree that standardized tests have limitations, but it is not fair to dismiss standardized testing for that reason alone. After all, other institutional assessment methods suffer from limitations that standardized tests do not. For example, student surveys (e.g., asking students how much they learned in college) do not measure learning directly. Alumni surveys, which ask graduates about how well their education prepared them for employment, do not provide up-to-date information about educational programs that might be improved nor is it clear whether positive outcomes reported on these surveys were the result of college education or something that occurred in the years following graduation. Portfolios provide direct evidence of students' academic development, but they cannot facilitate fair comparisons between students or institutions because they lack adjustments for task difficulty (even if a standardized evaluation rubric is employed).

By focusing on what tests do not measure, critics fail to give appropriate consideration to the relative importance of particular general outcomes measured by available standardized tests. These general outcomes transcend academic programs and institutions and therefore have the potential to be measured using standardized tests. Indeed, many postsecondary institutions have general education goals that could be measured using standardized tests. As a recent survey revealed, "A large majority of AAC&U member institutions (78%) say that they have a common set of intended learning outcomes for *all* their undergraduate students...The skills most widely addressed are writing, critical thinking, quantitative reasoning, and oral communication skills..."<sup>30</sup> The importance of these skills is also widely recognized outside of higher education. Leaders in business and government regard critical thinking, writing, and other so-called "higher-order" skills as essential for accessing and analyzing the information needed to address

**Draft: Please do not cite without permission from the author**

the complex, non-routine challenges facing workers and citizens in the twenty first century.<sup>31</sup> Along these lines, a full 95% of employers (and more than 90% of recent graduates) affirmed the importance of providing a four-year college education that includes domain-specific knowledge and skills as well as “intellectual and practical skills that span all areas of study, such as communication, analytical, and problem-solving skills, and a demonstrated ability to apply knowledge and skills in real-world settings.”<sup>32</sup> In the same survey, nearly 75% of employers said that colleges should place greater emphasis on teaching critical thinking and communication skills. By implication, employers feel that college graduates are deficient in these domains.

So, postsecondary schools might consider measuring general higher-order skills as part of an institutional assessment program because most schools uphold teaching these skills as a major goal of undergraduate education and the beneficiaries of college education are particularly concerned about these skills. Though, it must still be determined whether tests can adequately measure general higher order skills. At present, standardized testing seems like a sensible approach to measuring critical thinking and writing skills in particular because past research indicates that college education has a positive effect on these outcomes and that certain instructional practices are likely to promote the development of these skills.<sup>33</sup> Thus, it has been demonstrated that growth in these skills is measurable, and guidance for improving instruction on these skills is available to underperforming schools.

Furthermore, unlike some desirable general outcomes of higher education (e.g., leadership and preparedness for working in a global society), critical thinking and writing have operational definitions that are generally agreed upon,<sup>34</sup> and this allows for the development of tests that elicit responses many would accept as evidence of students’ abilities in those domains. In the case of critical thinking, tests should be designed to elicit evidence that students can

**Draft: Please do not cite without permission from the author**

identify the strengths and weaknesses of multiple perspectives on an issue, recognize connected and conflicting information, evaluate evidence as credible or unreliable, detect flaws in logic and questionable assumptions, acknowledge the need for additional information in the face of uncertainty, and weigh evidence from multiple sources to make a decision. With regard to writing, tests should elicit evidence that students can communicate ideas clearly and accurately as part of an organized and logically cohesive argument and control the elements of standard written English. Note that available standardized tests do not measure dispositions to think critically nor can they reasonably evaluate students' skills related to drafting and revising text.

Of currently available standardized tests, CAAP, MAPP, and CLA all claim to measure critical thinking and writing skills. CAAP and MAPP offer multiple-choice tests and optional essays for measuring critical thinking and writing. The CLA comprises open-ended problems that require students to evaluate, analyze, and synthesize information in order to articulate a position, critique an argument, or make an informed decision. Recently, ACT, ETS, and CAE collaborated on a test validity study supported by the Fund for the Improvement of Postsecondary Education.<sup>35</sup> In this study, college freshmen and seniors from a diverse sample of schools each took some combination of CAAP, MAPP, and CLA test modules. Results revealed that institutional average scores were highly reliable for all tests (typically greater than 0.85) even with sample sizes of less than 50 students per school and that institutional scores on all tests were highly correlated (typically greater than 0.90). The latter result suggests that schools with students possessing the skills needed to perform well on one test tend to have students with the skills needed to perform well on other tests. Consequently, it matters little which test is administered when ranking schools is the goal. Thus, choice of an assessment is likely to be

**Draft: Please do not cite without permission from the author**

guided by alignment with the learning goals of the institution, faculty acceptance, and practical considerations such as cost and ease of administration.

Institutional assessment focuses on the entire curriculum (even the whole university experience), not on the specific details of the learning environment. This broad scope explains why results from standardized tests of general college outcomes cannot diagnose the causes of academic strengths and weaknesses. However, much like blood pressure and cholesterol levels serve as useful indicators of illness in medicine (but not of the exact underlying cause for the illness), standardized test scores indicate the presence (or absence) of potential problems in academic domains. Even outspoken critic Trudy Banta acknowledges that standardized tests of general skills are “appropriate for giving individual students and faculty ideas about strengths and weaknesses and areas of curriculum and instruction to improve...”<sup>36</sup> If school administrators determine that test results are unacceptable, additional research is called for to determine which aspects of the learning environment might be improved. Alternatively, when test results are good, administrators might want to research what the school is doing right. One common approach involves interpreting test scores in light of results from student surveys such as the National Survey of Student Engagement (NSSE), a self-report measure that asks students about their behaviors inside and outside of the classroom that are known to be associated with academic and personal growth.

Indeed, standardized test results (along with data from complementary research) have informed reevaluations of general education programs at many schools.<sup>37</sup> This includes reform efforts in the University of Texas system, where CLA results were used to identify and address deficiencies in writing skills. At Stonehill College, CLA and NSSE results led to questioning of the level of academic challenge. Subsequent discussions led to modifications of the course-credit

**Draft: Please do not cite without permission from the author**

model to provide additional opportunities to take challenging courses and to the development of first-year seminars that include a focus on critical thinking and writing skills. At Seton Hill University, disappointing CLA scores and NSSE results suggesting low levels of academic challenge contributed to the decision to implement bi-weekly professional development on teaching critical thinking as well as additional requirements for writing in each major. In another example, CLA results at Barton College revealed performance above statistical expectations, but below personal expectations.<sup>38</sup> As a result, Barton revised its general education curriculum to require courses that specifically target critical thinking and written communication skills (4 courses each) and included CLA score targets as part of their reaccreditation “Quality Enhancement Plan.”

Furthermore, if standardized tests set a good example for assessment, these tests can reinforce classroom teaching and assessment of higher-order skills. For instance, given the great emphasis on teaching critical thinking in college and university missions, instructors should infuse classroom assessments (and other instructional activities) with opportunities for students to practice and demonstrate critical thinking. To this end, critical thinking tasks like those included in CAAP, MAPP, and the CLA can be tailored to the content and goals of courses in a broad spectrum of academic domains. At some schools, CLA-like performance tasks have been administered to measure general education outcomes as well as outcomes in chemistry, quantitative methods, and linguistics courses.<sup>39</sup>

Finally, if colleges and universities should be measuring higher-order skills, if it is possible to measure them with standardized tests, if test results serve as a useful indicator of institutional strengths and weaknesses, and if the tests encourage classroom teaching and assessment that is aligned with general education goals, then schools should be administering

**Draft: Please do not cite without permission from the author**

standardized tests of general college outcomes. However, as noted earlier, institutional test scores tend to be highly correlated with educational inputs, so test results, while providing useful information about the absolute level of student achievement, do not afford non-selective institutions the opportunity to demonstrate their educational efficacy for accountability purposes. To address this difficulty, the following section examines the potential for using value-added modeling as a tool for estimating a school's contribution to student learning after controlling for differences in entering academic ability.

**Value-Added Modeling**

Highly selective institutions typically rank favorably when compared with other schools using average test scores, but research suggests that an institution's selectivity provides little information about its actual contributions to learning over the course of college.<sup>40</sup> With institutional value-added modeling, each school's performance is compared to expectations established by the entering academic ability levels of their respective students. A non-selective school might perform poorly in an absolute sense, but it might obtain a high value-added score by performing much better than one would expect of a school admitting students of certain entering ability.

The dominant institutional value-added model currently used in higher education was first employed by CAE during the 2004-2005 CLA administration.<sup>41</sup> The basic idea behind this model is to compare the average difference in CLA scores between freshmen and seniors to the average difference one would expect based on their respective average entering academic ability levels (as measured by the SAT or ACT). When a school's average freshman-senior difference exceeds expectations, this indicates above expected learning gains at the school, and the school

obtains a high value-added score. This value-added approach has also been used recently by CAAP and MAPP for schools participating in VSA. This decision followed from research at ACT and ETS that demonstrated the potential for using scores from their respective assessments for estimating value added using this model.<sup>42</sup>

While value-added modeling may sound like an appealing way to evaluate educational quality, value-added scores, like all inferential statistics, are estimates of unknown values and therefore have inherent uncertainty that arises primarily from the fact that value-added scores might have come out differently if the schools tested different samples of students. The extent of this uncertainty is typically evaluated by estimating the *reliability* of the scores, which is indexed by a coefficient that ranges from 0 to 1. For the purposes of this chapter, high reliability coefficients indicate that value-added scores are precisely estimated (i.e., contain little random error) and that schools would receive very similar scores if they repeated the testing with different but similarly representative samples of students.

The major argument that critics levy against value-added modeling is that value-added scores are not nearly reliable enough to be interpreted validly as indicators of relative learning gains. Citing a review of literature, Trudy Banta reported the reliability of value-added scores to be approximately 0.10.<sup>43</sup> In an analysis of longitudinal data from students taking the multiple-choice version of COMP, Gary Pike estimated the reliability of gain scores as 0.14 and the reliability of residual scores as 0.17.<sup>44</sup> It must be noted, however, that these reliability coefficients are based on models that treat the student as the primary unit of analysis. Just as school means are more reliable than individual student scores, one might expect institutional value-added scores to be more reliable than student gain scores or residual scores.

**Draft: Please do not cite without permission from the author**

On the other hand, reliable institutional comparisons may simply be impossible because nearly all of the score variation is within schools on collegiate measures like NSSE, when between-school variation is required for institutional comparisons.<sup>45</sup> As further reason to distrust value-added modeling, critics also point to evidence that different value-added models lead to different results.<sup>46</sup> In that case, it would not be clear which result to trust. Several other criticisms of CLA value-added assessment not directly related to the statistical model (e.g., sampling issues, use of cross-sectional data, and concerns about student motivation) are addressed elsewhere.<sup>47</sup>

*Are Institutional Value-Added Scores Reliable?*

Tests scores with reliability of 0.80 or greater are commonly upheld as adequately reliable, but the true determination of adequacy depends heavily on the intended use of the test scores. If test scores are to be used in a high stakes context to make consequential decisions about students, teachers, or schools, reliability greater than 0.80 might be called for. On the other hand, reliability less than 0.80 would likely suffice if test results are used for a relatively low stakes purpose like identifying academic strengths and weaknesses to help improve teaching learning. In efforts to improve educational programs, it is also worth considering the relative reliability of various sources of information. As explained by Derek Bok, “The proper test for universities to apply is not whether their assessments meet the most rigorous scholarly standards but whether they can provide more reliable information than the hunches, random experiences, and personal opinions that current guide most faculty decisions about education.”<sup>48</sup>

As large-scale efforts to compute institutional value-added scores in higher education are fairly recent, there has only been one study published thus far providing the reliability of value-added scores generated by the original CLA value-added approach. In this study, Stephen Klein,

**Draft: Please do not cite without permission from the author**

Roger Benjamin, Richard Shavelson, and Roger Bolus, all staff members or consultants at CAE, made the case that value-added scores based on CLA results are adequately reliable.<sup>49</sup> They employed a novel split-sample approach to estimating reliability that involved randomly splitting available data from each class (freshmen and seniors) in each school, estimating separate value-added scores for two samples, and computing the correlation between them.

The reliability coefficient reported by Klein and his colleagues, 0.63, could be considered high enough to warrant interpretations of value-added scores as indicators of learning gains relative to expected for low stakes purposes. Klein's results were recently extended by correcting for the use of half-size samples and by computing the mean of 1,000 possible random splitting instead of just 1.<sup>50</sup> These analyses were carried out using data from 99 schools that administered the CLA in the 2006-2007 academic year and 154 schools from 2007-2008. Value-added reliability computed using these methods was 0.73 for the 2006-2007 data and 0.64 for the 2007-2008 data. These values provide higher (and more accurate) estimates of value-added score reliability, but room for improvement is still evident. One possible contributor to unreliability is that the original CLA value-added approach depends on difference scores (e.g., the difference between freshman and senior mean CLA scores), which are known to be less reliable than the scores from which they are derived.

This apparent unreliability may be manifested in low year-to-year consistency in value-added scores. Indeed, some schools express puzzlement about seemingly unrealistic swings in value-added scores across years. Of course, one should not expect value-added scores to be the same every year (e.g., due to programmatic changes, major differences in sampling methods, or measurement error), but one should not expect them to change radically either. Using the data from 71 schools participating in both CLA administrations, the correlation between value-added

**Draft: Please do not cite without permission from the author**

scores across years was only 0.32, which suggests that sizable year-to-year fluctuations in value-added scores were fairly common. For instance, Allegheny College's change from "at expected" to "below expected" value added coincided with a change from administering the test during the first few weeks of classes to administering the test during orientation (when students were "burned out, overwhelmed, and had too much going on").<sup>51</sup> This sort of variability in value-added scores, which is not always so easily explained, does not bode well for institutional assessment programs that seek to measure improvement over time.

Briefly, with regard to the concern over between- and within-school variance, data from recent CLA administrations reveals that approximately 20% of the variance in CLA scores is between schools (compared to less than 5% on some NSSE scales). As the percentage of between-school variance approaches zero, one would expect the reliability of between-school comparisons to approach zero. Given the results presented above, it seems that 20% between-school variance is sufficient to allow for reasonably reliable institutional comparisons.

In sum, critiques concerning the unreliability of institutional value-added scores are often overstated due to their orientation toward student-centered value-added models rather than models treating the school as the unit of analysis. Reliability analyses that treat the school as the unit of analysis reveal that institutional value-added scores generated by the original CLA value-added model are reliable enough to serve as indicators of relative learning gains, but not reliable enough to allow trustworthy institutional comparisons or to measure small changes in value added over time.

#### *An Alternative Value-Added Model*

CAE researchers evaluated other possible value-added models and identified one model as particularly promising because it improved upon the reliability of value-added scores

**Draft: Please do not cite without permission from the author**

(especially across years) and allowed for the calculation of institution-specific indicators of value-added precision. The lack of these indicators was a major drawback of the original CLA value-added approach because, for example, a school testing a very small number of students would have a less precise value-added estimate, but the original CLA value-added approach could not quantify the degree of uncertainty in that estimate. Standard errors signal to each campus the uncertainty in its value-added score and therefore facilitate honest interpretations of the significance of differences between schools.

The alternative value-added estimation approach employs a hierarchical linear model (HLM) with two levels of analysis: a student level for modeling within-school CLA score variation and a school level for estimating institutional value-added scores.<sup>52</sup> Rather than computing value-added based on difference scores (a possible source of unreliability), the new model works by comparing senior average CLA scores to expected CLA scores based on entering academic ability as measured by average SAT or ACT scores and average freshman CLA scores (serves as a control for selection effects not covered by the SAT or ACT). Although this model does not provide a direct measure of growth between freshman and senior year, it still works as a value-added model because, if one observes a group of seniors performing better than expected, it suggests that more learning took place at their school than at the typical school with comparable entering academic ability. Modeling within-school variance at the student level allows for computing standard errors of the value-added scores.

The same analyses described above were carried out on the same data using value-added scores derived from the HLM-based value-added model. Results indicate that value-added score reliability increased to 0.81 in 2006-2007 and 0.75 in 2007-2008 with the new model, and the correlation between value-added scores across years increased substantially from 0.32 to 0.58.

**Draft: Please do not cite without permission from the author**

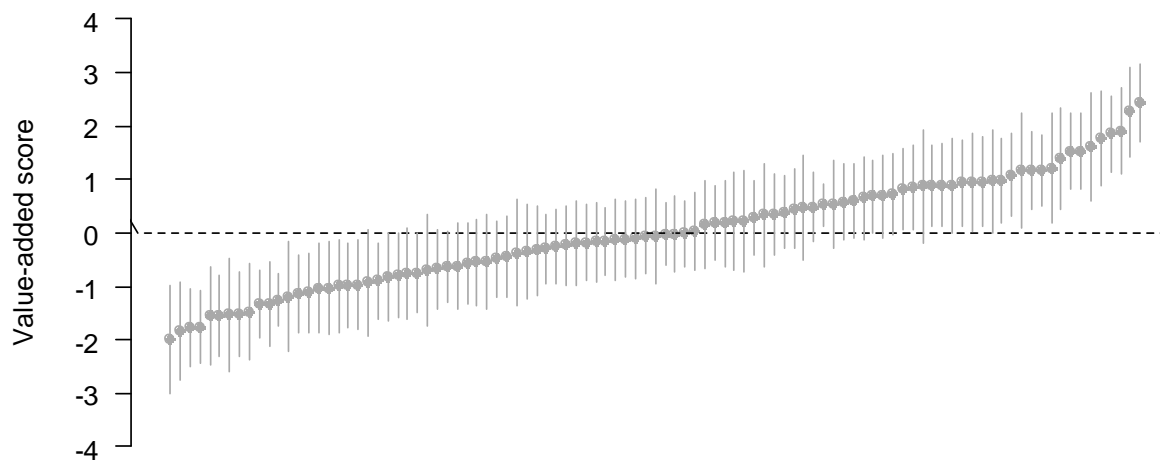
These results suggest that value-added scores based on the new model are more sensible for use in comparing institutions and have more realistic variation across years. Still, any differences between value-added scores should be evaluated in light of their standard errors (explained below).

To address concerns about different value-added models producing different results, correlations were computed between scores generated by the original CLA model and by the HLM-based model. The correlation was 0.80 in 2006-2007 and 0.72 in 2007-2008. These correlations reveal that the schools line up in similar but not identical ways based on the value-added scores derived from the two models. However, after disattenuating for unreliability (using the coefficients provided above), the correlation was 1.00 in both years (rounded down from values slightly higher than 1.00), which suggests that the value-added scores would be identical apart from measurement error.

Additional evidence from simulated CLA data supports this conclusion. CLA and SAT data for 200 simulated schools were generated using parameters from actual CLA schools (means, standard deviations, and covariance matrices). These data were used to estimate value-added scores under the condition that all schools tested all students (i.e., with no sampling error). When these full simulated data sets were used, the correlation between value-added scores was 1.00. Thus, different value-added models may produce different scores, but they may be estimating the same underlying construct. In other words, if all schools tested all students, it would not matter which value-added model was used because results would be identical. However, since financial and practical constraints prevent this from happening, it makes sense to choose the value-added model that provides more reliable scores for a given number of students

tested. For this reason, the new, HLM-based model will be used for the CLA starting in the 2009-2010 administration.

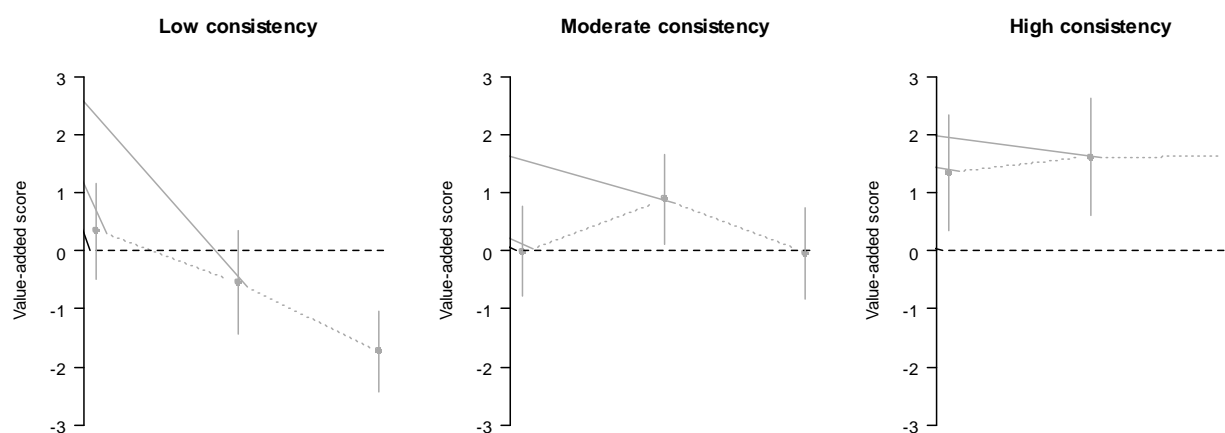
In addition to improvements in reliability, the HLM-based model provides a standard error for each school's value-added score, which may be used to compute 95% confidence intervals. By providing confidence intervals, schools get a realistic sense of the variability in value-added scores they could expect if testing was repeated with different students. The intervals also provide a sense of which differences in value-added scores could be considered significant. Figure 1 shows the value-added estimates and 95% confidence intervals for schools in 2006-2007 in order of increasing value-added (value-added scores are expressed in standard units). The influence of sample size on confidence interval size is noticeable in this figure (schools testing the most students have the smallest confidence intervals).



*Figure 1.* Value-added scores (shown as dots) and 95% confidence intervals (vertical lines) in 2006-2007.

To further investigate value-added stability over time, value-added scores and corresponding 95% confidence intervals were computed for 40 schools that participated in each of 3 recent CLA administrations (from 2005-2006 to 2007-2008). The ranges of their value-added scores were compared to the maximum of their 95% confidence intervals across the 3

administrations. Only 10% of schools had ranges that exceeded the size of their respective maximum 95% confidence intervals, which suggests that the confidence intervals provide a realistic indicator of random variation in value added scores. Figure 2 shows examples of value-added score consistency that is low (range greater than 1.5), moderate (range between 0.75 and 1.5), and high (range less than 0.75). Under the original value-added model, 40% of schools have low consistency, 38% have moderate consistency, and 23% have high consistency. With the new value-added model, a much higher percentage of schools have moderate rather than low consistency (25% low, 55% moderate, and 20% high).



*Figure 2.* Examples of low, moderate, and high value-added score consistency across 3 CLA administrations.

Since most schools have moderate to high consistency with the HLM-based model, value-added scores, despite their inherent uncertainty, appear to be capturing fairly stable characteristics of schools. This supports the interpretability of value-added scores as indicators of relative learning gains, but the degree of remaining uncertainty makes it difficult to reliably compare schools with similar value-added scores or to detect small increases in value-added

**Draft: Please do not cite without permission from the author**

scores due to programmatic improvements. As an additional caution, it seem advisable to examine value-added scores across multiple years or to cross-validate value-added results with other indicators of student learning to ensure that any consequential decisions are based on dependable information.

**Conclusions and Recommendations**

There are limitations to currently available standardized tests of general college outcomes, but there still are many reasons to administer them. Standardized tests may not measure every important outcome of college education, but they measure some especially salient ones like critical thinking and communication skills, which are viewed as essential for making contributions in the workplace and successfully navigating life in the twenty first century. Standardized tests results do not prescribe specific ways to improve learning, but they signal possible problems by providing indicators of students' academic strengths and weaknesses. Finally, despite the bad reputation that standardized tests have acquired among their critics, it is possible for standardized tests to reinforce institutional goals and to set a good example for classroom assessment.

With regard to value-added modeling, new evidence indicates that institutional value-added scores are significantly more reliable than critics have claimed. Furthermore, advances in institutional value-added modeling promise greater score reliability, more realistic consistency across years, and improved score interpretability while testing the same number of students. At this time, the statistical dependability of value-added scores is likely sufficient to allow for a rough ordering of schools by their contributions to student learning on outcomes measured by standardized tests. However, the degree of remaining statistical uncertainty precludes the

**Draft: Please do not cite without permission from the author**

possibility of measuring small changes over time or using value-added scores to guide consequential decisions based on comparisons of schools. That being the case, the principal utility of value-added scores is that they provide signals of educational quality and should therefore stimulate discussions about improvement and prompt additional research to fully determine the underlying reasons for academic strengths and weaknesses. According to Terrence Grimes, Vice President for Academic Affairs at Barton College, “The college’s work with the CLA proved to be a way to draw administrators, faculty and staff members, and even trustees into a frank, sometimes difficult but ultimately constructive conversation about how well Barton was actually educating its students.”<sup>53</sup>

In order to maximize the quality and interpretability of value-added scores, it is recommended that schools sample students randomly for testing if possible, meet sample size guidelines, and consider in-depth sampling of sub-units within institutions. This final recommendation is especially important for large schools where the most appropriate unit of analysis may be colleges within the university rather than the entire university.

As evidenced by examples provided above from the University of Texas, Stonehill College, Seton Hill College, and Barton College, standardized tests seem to be providing data that schools can act upon, but they do not offer a complete solution to the challenges of institutional assessment. As explained by Lee Shulman, President of the Carnegie Foundation for the Advancement of Teaching, “...nearly any goal of using the results of assessment for serious practical and policy guidelines should intentionally employ an array of instruments.”<sup>54</sup> For example, evaluation programs in the University of Texas system include five components: “persistence and graduation rates, license exam pass rates in critical fields, postgraduate experience, student experience, and student learning assessments.”<sup>55</sup> By comparing and

**Draft: Please do not cite without permission from the author**

contrasting evidence from a variety of sources, schools obtain a more complete picture of educational quality and help ensure that changes to general education programs are informed by robust information.

Standardized tests may also be supplemented with other direct measures of student learning such as electronic portfolios that allow students to upload and organize evidence of learning such as papers, presentations, artwork, videos, and audio recordings. It is unfortunate that standardized tests and portfolios are frequently pitted as arch nemeses (with proponents of each conveniently ignoring the shortcomings of their favored approach) because there is possibility for overlap. Technology exists that would allow for the inclusion of student responses to standardized open-ended assessment tasks in a portfolio with scores reported on a standard scale. This could benefit portfolios by adding a standardized component, which might improve the comparability of portfolios across students and schools.

Legislators take note that standardized testing may be imposed on schools, but this cannot guarantee that schools will put forth the effort and resources required to administer the test properly or that results will be put to good use (more likely than not, it will have the opposite effect). It is unfortunate that, at many schools, student outcomes assessment is mostly a matter of “going through the motions” for accreditation or state mandated accountability. Many faculty members have genuine misgivings about standardized testing, and some have gross misperceptions (e.g., that the CLA is a multiple-choice test). Negative sentiment among faculty is passed along to students, and, in some cases, students have been told “the test doesn’t matter.” This would likely undermine any possibility of obtaining results that could be interpreted validly.

In order for standardized testing and value-added assessment to gain widespread acceptance in higher education, it seems that any efforts to implement standardized tests

**Draft: Please do not cite without permission from the author**

(whether voluntary or mandated by legislation) must be accompanied by programs that engage faculty and help foster a culture of assessment and improvement. For any school considering a standardized test of general college outcomes, Trudy Banta offers sensible guidelines for involving faculty, staff, and students in the process.<sup>56</sup> Namely, schools should ensure that tests align with general educational goals of the institution, have faculty and students take the test and review their scores, conduct cognitive analyses of the test items (e.g., asking students to think aloud while taking the test), publicize the decision, and continually evaluate the usefulness of a test as it is administered.

Schools might take additional steps to establish buy-in from faculty for standardized testing through professional development programs like the *CLA in the Classroom Academy*, a two-day workshop designed to introduce instructors to teaching and assessment practices that align with general education goals. Through the course of the workshop, participants learn to develop CLA-like performance tasks for use in their classrooms and to apply scoring rubrics that provide students with formative feedback. Evidence from workshop evaluations reveals that some reluctant participants have developed newfound enthusiasm for student assessment at the classroom and institution levels, and several schools now embed performance tasks in their general education programs.

Despite efforts to create a culture of evidence and assessment, there will be holdouts at many schools who dismiss standardized test results and other findings from institutional research efforts. One way to overcome this roadblock would be to provide the tools and resources needed to develop local assessments that would allow faculty to demonstrate learning gains on the knowledge and skills they claim to teach and to carry out small-scale research on improving instruction. Faculty members are unlikely to persist in claiming to teach something without

**Draft: Please do not cite without permission from the author**

providing evidence of learning or to express concern about instructional quality without evaluating it. If many professors (or entire departments) rise to the challenge proposed here, their work could possibly stimulate interesting developments in classroom instruction and assessment.

In the near future, standardized test providers should carry out the research needed to supply additional evidence that value-added scores can be interpreted validly as indicators of learning gains relative to expected. Specifically, it has yet to be demonstrated that cross-sectional and longitudinal data collections would produce similar results. Additionally, the relationships among institutional value-added scores, other school-level indicators of student learning, and institutional characteristics should be examined. This research could contribute to the validation of value-added scores, and it might help identify student behaviors and institutional characteristics that are associated with low and high value-added scores.

As for test development, there appears to be demand for major-specific standardized assessments that incorporate critical thinking and writing skills. Faculty in certain departments may be more inclined to take results from such assessments seriously because they would call upon students to demonstrate higher-order skills and apply content knowledge as a historian, an engineer, a social scientist, or a businessperson to name a few possibilities. There is also room for standardized tests of non-cognitive higher-order outcomes of higher education like personal and social responsibility, creativity, global competency, leadership, teamwork skills, and others. However, substantial research is called for in order to establish consensus definitions of these constructs and to demonstrate that they can be measured reliably and efficiently.

To conclude, standardized testing of general college outcomes and the value-added scores generated by such testing are valuable components of institutional assessment programs. Given their degree of statistical dependability, value-added scores serve as indicators of learning gains

**Draft: Please do not cite without permission from the author**

relative to expected, and they may be employed to roughly order schools based on their contributions to the development of skills measured by currently available standardized tests. However, they cannot presently provide dependable comparisons between schools with similar value-added scores. In light of this, policymakers and leaders in higher education should focus their attention on the original (and arguably most important) purposes of administering standardized tests of general college outcomes: indicating areas of possible weakness in general education programs and stimulating discussions about improving instruction. A large number of schools have already taken advantage of the opportunities afforded by currently available tests. As non-participating schools slowly come to recognize the positive changes that have resulted from concerted efforts to measure and improve learning using standardized tests, these measures should gain much wider acceptance and schools will take the steps necessary to ensure that results are put to good use.

---

<sup>1</sup> U.S. Department of Education, *A Test of Leadership: Charting the Future of U.S. Higher Education* (Washington, DC: 2006), p. 14.

<sup>2</sup> Business-Higher Education Forum, *Public Accountability for Student Learning in Higher Education: Issues and Options* (Washington, DC: American Council on Education, 2004).

<sup>3</sup> Association of American Colleges and Universities and the Council for Higher Education Accreditation, *New Leadership for Student Learning and Accountability: A Statement of Principles, Commitments to Action* (Washington, DC: AAC&U and CHEA, 2008), p. 1.

<sup>4</sup> National Research Council, Board on Testing and Assessment, *Lessons Learned About Testing: Ten Years of Work at the National Research Council* (Washington, DC: National Research Council, 2007).

<sup>5</sup> Thomas Webster, "A Principal Component Analysis of the U.S. News & World Report Tier Rankings of Colleges and Universities," *Economics of Education Review* 20, no. 3 (June 2001).

<sup>6</sup> Stephen Klein, George Kuh, Marc Chun, Laura Hamilton and Richard Shavelson, "An Approach to Measuring Cognitive Outcomes across Higher Education Institutions," *Research in Higher Education* 46, no. 3 (May 2005).

<sup>7</sup> William Bowen, Matthew Chingos and Michael McPherson, *Crossing the Finish Line: Completing College at America's Public Universities* (Princeton, NJ: Princeton University Press, 2009).

<sup>8</sup> Robert Lissitz (Ed.), *Value Added Models in Education: Theory and Applications* (Maple Grove, MN: JAM Press, 2005).

<sup>9</sup> American Association of State Colleges and Universities, "Value-Added Assessment: Accountability's New Frontier," *Perspectives*, (Spring 2006).

<sup>10</sup> Peter Ewell, "An Emerging Scholarship: A Brief History of Assessment" in *Building a Scholarship of Assessment*, eds. Trudy Banta and Associates (San Francisco: Jossey-Bass, 2002). C. Robert Pace, *Measuring Outcomes of College: Fifty Years of Findings and Recommendations for the Future* (San Francisco: Jossey-Bass, 1979). Richard Shavelson, *A Brief History of Student Learning Assessment: How We Got Where We Are and a Proposal for Where to Go Next* (Washington, DC: Association of American Colleges and Universities, 2007).

- 
- <sup>11</sup> William Learned and Ben Wood, *The Student and His Knowledge: A Report to the Carnegie Foundation on the Results of the High School and College Examination of 1928, 1930, and 1932* (Boston: The Merrymount Press, 1938), p. 371.
- <sup>12</sup> Reuben Frodin, "Very Simple but Thoroughgoing" in *The Idea and Practice of General Education: An Account of the College of the University of Chicago by Present and Former Members of the Faculty*, ed. F. Champion Ward (Chicago: The University of Chicago Press, 1950).
- <sup>13</sup> Executive Committee of the Cooperative Study in General Education, *Cooperation in General Education* (Washington, DC: American Council on Education, 1947).
- <sup>14</sup> Richard Shavelson (2007).
- <sup>15</sup> Richard Shavelson (2007).
- <sup>16</sup> ACT, *College Outcome Measures Project Revised Summary Report of Research and Development 1976-1980* (Iowa City, IA: ACT, 1981). Ernest Pascarella and Patrick Terenzini, *How College Affects Students: A Third Decade of Research* (San Francisco: Jossey-Bass, 2005). Laura Underwood, Barbara Maes, Lisa Alstadt and Michael Boivin, "Evaluating Changes in Social Attitudes, Character Traits, and Liberal-Arts Abilities During a Four-Year Program at a Christian College," *Research on Christian Higher Education* 3, (1996).
- <sup>17</sup> Peter Ewell (2002).
- <sup>18</sup> Ohio University Office of Institutional Research, *General Education Outcomes: The College Outcomes Measures Program (COMP) at Ohio University 1981-1985* (Athens, OH: Ohio University, 1996).
- <sup>19</sup> Peter Ewell, "Assessment and Accountability in America Today: Background and Context" in *Assessing and Accounting for Student Learning: Beyond the Spellings Commission: New Directions for Institutional Research, Assessment Supplement 2007*, eds. Victor Borden and Gary Pike (San Francisco: Jossey-Bass, 2008).
- <sup>20</sup> Business-Higher Education Forum (2004). State Higher Education Executive Officers, *Accountability for Better Results: A National Imperative for Higher Education* (Boulder: State Higher Education Executive Officers, 2005). U.S. Department of Education (2006).
- <sup>21</sup> Stephen Klein, Roger Benjamin, Richard Shavelson, and Roger Bolus, "The Collegiate Learning Assessment: Facts and Fantasies," *Evaluation Review* 31, no. 5 (October 2007).
- <sup>22</sup> Hart Research Associates, *Learning and Assessment: Trends in Undergraduate Education - A Survey among Members of the Association of American Colleges and Universities* (Washington, DC: Hart Research Associates, 2009), p. 9.
- <sup>23</sup> Council of Independent Colleges, *Evidence of Learning: Applying the Collegiate Learning Assessment to Improve Teaching and Learning in the Liberal Arts College Experience* (Washington, DC: Council of Independent Colleges, 2008).
- <sup>24</sup> Council of Independent Colleges, *CIC/CLA Consortium Resources* (Washington, DC: Council of Independent Colleges, 2009). Available online at [http://www.cic.edu/projects\\_services/coops/cla\\_resources/index.html](http://www.cic.edu/projects_services/coops/cla_resources/index.html)
- <sup>25</sup> Peter McPherson and David Shulenburg, *Toward a Voluntary System of Accountability (VSA) for Public Universities and Colleges* (Washington, DC: National Association of State Universities and Land-Grant Colleges, 2006).
- <sup>26</sup> Christine Keller and John Hammang, "The Voluntary System of Accountability for Accountability and Institutional Assessment" in *Assessing and Accounting for Student Learning: Beyond the Spellings Commission: New Directions for Institutional Research, Assessment Supplement 2007*, eds. Victor Borden and Gary Pike (San Francisco: Jossey-Bass, 2008), p. 45.
- <sup>27</sup> Trudy Banta and Gary Pike, "Revisiting the Blind Alley of Value Added," *Assessment Update* 19, no. 1 (January-February 2007), p. 15.
- <sup>28</sup> AAC&U and CHEA Association of American Colleges and Universities and the Council for Higher Education Accreditation (2008), p. 5.
- <sup>29</sup> Derek Bok, *Our Underachieving Colleges: A Candid Look at How Much Students Learn and Why They Should Be Learning More* (Princeton: Princeton University Press, 2006), p. 251.
- <sup>30</sup> Hart Research Associates (2009), p. 2.
- <sup>31</sup> The New Commission on the Skills of the American Workforce, *Tough Choices or Tough Times* (Washington, DC: National Center on Education and the Economy, 2006). The Secretary's Commission On Achieving Necessary Skills, *What Work Requires of Schools: A Scans Report for America 2000* (Washington, DC: U.S. Department of Labor, 1991).
- <sup>32</sup> Hart Research Associates (2006).
- <sup>33</sup> Ernest Pascarella and Patrick Terenzini (2005). Elizabeth Jones, Steven Hoffman, Lynn Moore, Gary Ratcliff, Stacy Tibbetts, and Benjamin Click, *National Assessment of College Student Learning: Identifying College*

*Graduates Essential Skills in Writing, Speech and Listening, and Critical Thinking* (Washington, DC: National Center for Education Statistics, 1995).

<sup>34</sup> Elizabeth Jones et al. (1995).

<sup>35</sup> Stephen Klein, Ou Lydia Liu, and James Sconing, *Test Validity Study (TVS) Report* (2009). Available online at [http://www.voluntarysystem.org/docs/reports/TVSReport\\_Final.pdf](http://www.voluntarysystem.org/docs/reports/TVSReport_Final.pdf)

<sup>36</sup> Trudy Banta, "Editor's Notes: Trying to Clothe the Emperor," *Assessment Update* 20, no. 2 (March-April 2008), p. 4.

<sup>37</sup> Council of Independent Colleges, (2008). Richard Ekman and Stephen Pelletier, "Reevaluating Learning Assessment," *Change* 40, no. 4 (July-August 2008). Reyes and Rincon, (2008).

<sup>38</sup> Alan Lane and Kevin Pennington, *General Education, the QEP and the CLA: Barton College*, Presented at the CIC/CLA Consortium Summer Meeting (Jersey City, NJ, 2009).

<sup>39</sup> Kristy Miller, Gerald Kruse, Christopher LeCluyse and Joel Frederickson, *CLA Performance Tasks*, Presented at the CIC/CLA Consortium Summer Meeting (Jersey City, NJ, 2009).

<sup>40</sup> Ernest Pascarella and Patrick Terenzini (2005).

<sup>41</sup> Stephen Klein et al. (2007).

<sup>42</sup> ACT, *Voluntary System of Accountability Learning Gains Methodology* (Iowa City, IA: ACT, 2009). Available online at [http://www.voluntarysystem.org/docs/cp/ACTReport\\_LearningGainsMethodology.pdf](http://www.voluntarysystem.org/docs/cp/ACTReport_LearningGainsMethodology.pdf). Ou Lydia Liu, *Measuring Learning Outcomes in Higher Education Using the Measure of Academic Proficiency and Progress (MAPP)* (ETS RR-08-47) (Princeton, NJ: ETS, 2008). Available online at

<http://www.voluntarysystem.org/docs/cp/RR-08-47MeasuringLearningOutcomesUsingMAPP.pdf>

<sup>43</sup> Trudy Banta (2008).

<sup>44</sup> Gary Pike, *Lies, Damn Lies, and Statistics Revisited: A Comparison of Three Methods of Representing Change*, Paper presented at the Annual Forum of the Association for Institutional Research (San Francisco, 1992).

<sup>45</sup> George Kuh, *Director's Message - Engaged Learning: Fostering Success for All Students* (Bloomington, IN: National Survey of Student Engagement, 2006).

<sup>46</sup> Trudy Banta and Gary Pike (2007).

<sup>47</sup> Stephen Klein, David Freedman, Richard Shavelson, and Roger Bolus, "Assessing School Effectiveness," *Evaluation Review* 32, no. 6 (December 2008).

<sup>48</sup> Derek Bok (2006), p. 320.

<sup>49</sup> Stephen Klein et al. (2007).

<sup>50</sup> Jeffrey Steedle, *Advancing Institutional Value-Added Score Estimation (draft)* (New York: Council for Aid to Education, 2009). Available online at

<http://www.collegiatelearningassessment.org/files/AdvancingInstlValueAdded.pdf>

<sup>51</sup> Doug Lederman, "Private Colleges, Serious About Assessment," *Inside Higher Ed*, August 4, 2008. Available online at <http://www.insidehighered.com/news/2008/08/04/cla>

<sup>52</sup> Jeffrey Steedle (2009).

<sup>53</sup> Council of Independent Colleges (2008), p. 18.

<sup>54</sup> Lee Shulman, *Principles for the Uses of Assessment in Policy and Practice: President's Report to the Board of Trustees of the Carnegie Foundation for the Advancement of Teaching* (Stanford, CA: Carnegie Foundation for the Advancement of Teaching, 2006). Available online at

[http://www.teaglefoundation.org/learning/pdf/2006\\_shulman\\_assessment.pdf](http://www.teaglefoundation.org/learning/pdf/2006_shulman_assessment.pdf)

<sup>55</sup> Pedro Reyes and Roberta Rincon (2008), p. 51.

<sup>56</sup> Trudy Banta (2008).