RISING
TO THE CHALLENGE

Meaningful Assessment
of Student Learning

FIPSE

RISING
TO THE CHALLENGE

Meaningful Assessment
of Student Learning

Association of American Colleges and Universities
American Association of State Colleges and Universities
Association for Public and Land-grant Universities

Contents

Overview

T hat higher education makes an enormous positive difference in the lives of its students and produces benefits that elevate all of society was once an unchallenged assumption. The thought of measuring the amount of "learning" a college education generated simply did not occur to educators, legislators, or citizens because of the depth of this conviction about higher education.

The assessment movement that began in the 1980s represented the first effort to determine whether individual courses or portions of the curriculum accomplished the ends intended for them. This movement was largely internally motivated by educators who saw enough variation in the quality of education to suggest that some efforts were better than others. But the one-two punch that really got higher education's attention came in 2006 with the release of *A Test of Leadership*, the final report of Commission on the Future of Higher Education, and the publication of Derek Bok's book *Our Underachieving Colleges*.

A Test of Leadership directly challenged the assumptions about the outcomes of higher education's efforts and challenged us to test to see what value higher education added for its students. Bok's book, written by the ultimate insider, argued that improvement of higher education depended on measurement of outcomes.

Higher education responded to the call for measurement. The Association for Public and Land-grant Universities (APLU) teamed with the American Association of State Colleges and Universities (AASCU) to develop with its members the Voluntary System of Accountability (VSA), a system that included measurement of outcomes. The Association of American Colleges and Universities

(AAC&U) emphasized its LEAP program, focusing on the elements that matter in liberal education.

Our associations are firmly committed to improving higher education and to using measurement of outcomes as a tool in the effort. For that reason we jointly applied for a FIPSE grant that would help us perfect the art of measurement. The work done with that grant is described in the following pages. AASCU used its portion of the grant funds to develop a tool to measure the non-cognitive outcomes of learning; AAC&U used its portion to develop the art of applying rubrics to portfolios of student work to measure learning outcomes; and APLU used its funds to determine whether three standardized tests of learning outcomes used in the VSA measured essentially the same dimensions of learning.

We are grateful to FIPSE for funding this effort and express thanks to our member universities and colleges that participated in the research described here. We are pleased to present the results of that research to the higher-education community.

Carol Geary Schneider
President
AAC&U

Muriel A. Howard
President
AASCU

M. Peter McPherson
President
APLU

Executive Summary

Rising to the Challenge: Meaningful Assessment of Student Learning was envisioned in response to a 2007 request for proposals from the U.S. Department of Education's Fund for Improvement of Post Secondary Education (FIPSE). FIPSE called for national, consortial contributions to improving our knowledge and abilities to assess student learning for purposes of accountability and improvement. The Association of American Colleges and Universities (AAC&U), the American Association of State Colleges and Universities (AASCU), and the Association of Public and Land-grant Universities (APLU), collaboratively proposed three complementary projects to expand our understanding of the challenges and opportunities for assessing student learning.

The *Rising to the Challenge* project focused on refining our understanding of the most common, standardized instruments for measuring student learning and also on developing new tools and approaches for measuring and reporting student learning in a broad array of important areas of learning. The project involved the development of a new student survey, under the direction of AASCU, for areas of student learning that lack multiple measurement instruments—participation in civic engagement, preparation for success in the workplace, and acquisition of global skills. A second component, under the direction of AAC&U, involved the development of a set of nationally benchmarked rubrics articulating expected performance levels for 15 essential learning outcomes that can be used to assess student learning over time. The final part of the project involved a validity study, under the direction of APLU, of the three major standardized tests of student learning public universities are required to use to participate in the Voluntary System of Accountability's *College Portrait* Web reporting too. The three are the Measure of

Academic Proficiency and Progress (MAPP) (now renamed the ETS Proficiency Profile), ACT's Collegiate Assessment of Academic Proficiency (CAAP), and CAE's Collegiate Learning Assessment (CLA).

Each component of the project focused on institutional-level analysis rather than individual student assessment. Yet each component provides a foundation for examining or placing programmatic or individual assessments and information on student learning within a broader understanding or framework. Already, hundreds of faculty and campuses are drawing upon the findings and products of this project and are engaged in moving assessment of student learning forward on their respective campuses.

One of the fundamentals of good assessment practice is the need for multiple measures of student learning and success. Good assessment practice also supports measures that can help faculty, students, and others evaluate learning for both formative and summative purposes. The *Rising to the Challenge* project provides additional tools, information, and approaches for campuses to use to enhance their processes, practices, and reporting of student learning by creating a deeper understanding of the measures currently available, as well as two new measures.

This report begins with a description of the *Degrees of Preparation* project that developed a new student survey that will help campuses begin to measure learning outcomes affecting the public good. Too often student engagement with civic and political life, the skills and abilities associated with success in the workplace, and the acquisition of global knowledge and skills are neglected in the work on learning outcomes. Through extensive testing with students on all types of campuses across the country, the student-survey questions were developed and refined under the guidance of a national advisory panel and were field tested with over 3,000 students to establish clarity and verisimilitude.

The *Degrees of Preparation* survey provides not only data about student participation in, and perceptions of, their learning, but also asks students to share, through open-ended responses, their experiences with work and civic engagement. As a result, both quantitative and qualitative information is gathered for a multi-faceted data set on student learning in these important outcome areas.

The *VALUE* (Valid Assessment of Learning in Undergraduate Education) project addressed the question of whether there is a shared set of expectations among faculty about what student learning looks like for a broad range of outcomes. Over 100 faculty and staff members from every type of higher-education institution in the country were engaged in reviewing existing campus rubrics for 15 essential learning outcomes and in analyzing the rubrics for common, core criteria or elements of learning for each outcome. The project also called on the expertise and knowledge of experts in various academic fields to draft rubrics containing the dimensions considered essential or core to learning. Over 100 campuses field-tested the draft rubrics with either e-portfolios of student work or smaller samples of such work, to establish the rubrics' reliability and validity in measuring student learning in the one or more specific learning outcomes. Three rounds of rubric drafting, testing, re-drafting and testing again were conducted and then a national panel reviewed the work. This resulted in a final set of 15 rubrics that campuses can use as a national standard for learning at progressively more sophisticated levels as students move through and among our undergraduate institutions. The rubrics were found by faculty at all types of higher education institutions to be reliable and valid standards for assessing the quality of student learning.

In the third and final component of the project, a *Test Validity Study* (TVA) was conducted of the three learning-outcomes tests identified for use in the Voluntary System of Accountability (VSA)-The study addressed four questions about the tests: (1)

What is the reliability of school-level scores of different measures of writing and critical thinking? (2) To what degree do different measures designed to assess the same construct (such as critical thinking) correlate with each other as compared to tests that are designed to assess other constructs (such as reading)? (3) Is the average difference in mean scores (effect size) between freshmen and seniors similar across the different measures of the same construct? and (4) Do the scores on tests that use different response modes (such as essay or multiple choice) to assess a given competency (such as writing) correlate more highly with each other than they do with scores on tests that use the same response mode but assess different constructs? Through a carefully determined test matrix, test combinations were administered to 1,100 students at 13 colleges and universities. The results indicated that when the analysis was conducted at the campus level, all the tests ordered schools similarly, regardless of which constructs they were designed to measure or which response format was used.

What follows are brief summaries and descriptions of the three components of the *Rising to the Challenge: Meaningful Assessment of Student Learning* project. More complete details can be found on the respective associations' Web sites. The overall project has provided useful information about the opportunities and limitations of the most commonly used standardized tests of student learning; a new tool for gathering information about student learning in outcome areas previously ignored because of the lack of good, validated tools; and a new approach to assessing student learning that allows campuses to place their faculties' judgments of student performance within a nationally shared articulation of learning standards validated across institutional types.

In sum, the demand for information on the assessment of student learning has been enhanced in multiple ways through this collaborative project. The calls for assessment and accountability that prompted the current projects now have fuller answers. Higher education's multiple stakeholders can examine the measurement tools being used by campuses with a better understanding of what they provide; they can examine the articulated standards or expectations that faculty use to judge student learning quality and determine if they make sense; and they can understand more about the full range of student-learning outcomes that employers, community leaders, and colleges say our students need to be successful students and citizens in a global society. Although it was not an original motivation for the studies, perhaps one of the most valuable results of these projects is the richer understanding that students may gain for understanding and judging their own learning.

Readers are encouraged to contact the authors or to visit the associational web sites for more in-depth information about each of the projects.

Beginning to Measure Learning Outcomes Affecting the Public Good

John M. Hammang

Goals

AASCU's *Degrees of Preparation* project was designed to develop a student survey capable of measuring students' increasing preparation for participation in civic engagement, preparation for success in the workplace, and acquisition of global skills. It was developed as an institutional accountability measure and, as such, the survey's primary unit of analysis is the institution.

Project Background

U.S. Secretary of Education Margaret Spellings' *Commission of the Future of Higher Education* provided considerable focus on the lack of higher education's ability to report on learning outcomes. While content learning is an important outcome of a higher education, there are other outcomes also attributable to a higher education. Some very important outcomes affect the public good. They have long been recognized in the higher-education community, but there were no appropriate instruments to systematically and comprehensively report on these outcomes at an institutional level. This became very evident after a diligent search for such instruments during the development of the Voluntary System of Accountability. A technical work group focused on these issues failed to identify any such instrument. Secretary Spellings' decision to focus a major FIPSE grant on the development of accountability tools provided the opportunity to undertake development of such an instrument.

John Hammang is Director, Special Projects and Development at AASCU.

Project Organization

Survey development was organized around three sets of activities: identification of appropriate public-good learning outcomes related to an undergraduate experience, development of a survey instrument to measure changes between incoming freshmen and about-to-graduate seniors, and field testing the resulting instrument at a wide array of public and independent higher-education institutions. There is a detailed workflow chart in Appendix A.

Degrees of Preparation focuses on three high-profile areas of student growth:

■ Acquisition of skills for success in a global community
■ Preparation for success in the workplace
■ Preparation for participation in civic engagement

A panel of subject-matter experts in these three areas provided focus to the survey by identifying aspects of these issues where change could be defined and assurance that the areas of focus were relevant to various stakeholder communities served by college graduates. These discussions yielded an important realization: There is a significant overlap in the skills needed to prepare students for success in the workplace and for civic engagement following graduation. This will make correlation analysis important to see if student gains over time match changes in these skills.

The work of the experts panel became source material for drafting the survey items by a validation team. Members of the validation team were selected for their experience in developing, validating, and analyzing social-science survey instruments. The validation team created response scales, drafted survey items, conducted focus-group sessions with students, and conducted one-on-one cognitive interviews[1] with students. The student interactions were used to test whether the instrument was understood by students in the way intended by the item drafters and whether the scales

were functioning as intended. Revisions of the survey followed each of these procedures.

The survey instrument was field tested by 14 public and private institutions of higher education. The participating institutions spanned the geography of the United States, ranging from very large to very small. They represented both selective and non-selective admission policies and included both urban and rural campuses. Each of the field-test institutions conducted a full administration of the survey. This included submitting the survey instrument to the human-subjects review process (i.e., typically an institutional review board and in some cases for Web accessibility review). During the field tests 3,823 students completed the survey, and 557 students partially completed the survey.

Finally, post-field test statistical analyses were conducted to identify whether survey items yielded a useful dispersion of responses, whether related items were correlated as expected, and whether identifiable distinctions could be made between the responses of incoming freshmen and about-to-graduate seniors.

Project Methodology

The *Degrees of Preparation* survey is designed to be administered to a randomly drawn cross-sectional sample of first-time freshmen and about-to-graduate seniors (those who have earned 100 or more credit hours toward a baccalaureate degree). It measures and reports changes in preparation between the freshman and senior cohorts in the sample. The survey is administered online and takes about 15 minutes to complete. The survey instrument can be viewed at aascu.org/accountability/survey/?u=1.

In developing *Degrees of Preparation* as an instrument to measure change, we relied heavily on the concept that, going forward, students are more likely to do what they have already done and about which they have acquired some reasonable sense of personal competence. To focus on this concept, scales were

developed that allowed students to report how often they engaged in a particular behavior, to indicate how important various sources of information are to them, and to report their sense of personal competence in doing various activities. The survey designers excluded any report of competence that was accompanied by a report that the activity had not been engaged in by the student. In addition, survey items sequence skill-achievement queries in a Guttman-like scale that progresses from simple to increasingly complex skill accomplishments. This eases analysis of the results and also provides a mechanism to detect non-cooperative, random-answer patterns by survey takers.

Items useful for measuring preparation for civic engagement cover a swath of queries that collect information about:

- Sources of information
- Relative importance attached to various sources of information
- Political involvement
- Group skills *(items need revision)*
- Beliefs about community *(items need revision)*
- Helping others *(includes duration and intensity items, as well as inquiry about most meaningful experiences)*
- Critical thinking and communication skills
- Civic agency skills

Another focus of the development effort centered around two imperatives. The first is to begin to systematically and comprehensively collect information about learning outcomes that matter to community stakeholders and, secondly, to provide institutional leaders with student narrative reports of important experiences that made up part of their baccalaureate experience. In taking this approach, the survey developers hope to create a viable and robust platform that will allow institutions to make verifiable and credible claims about the public good that comes from higher education. The private good achieved through higher education has dominated community and policy discourse and has resulted in a long-term disinvestment in higher education

because of the evident private good that is achieved. This disinvestment overlooks the public good that is simultaneously achieved, and the survey developers hope to bring about a more balanced consideration of the public necessity of supporting higher education.

Student narratives of work and community experiences related to their baccalaureate experience are solicited during the course of the survey. Field-test results indicate a substantial willingness by survey takers to provide this qualitative information. As a practical matter, these narratives offer institutional leaders an opportunity to illustrate the quantitative information collected about these public good experiences using student voices and to provide a source of memorable stories that can underscore that message with stakeholder groups.

Results

Analysis of the data set developed by the field test of the survey instrument yielded a number of findings. Statistical treatment of the data set was designed to test whether the scales and items employed adequately differentiate among differing student experiences. In addition, the tests checked to see if the instrument would reveal important differences between the responses of freshman and seniors.

A large set of items in the survey used both a *frequency* scale (i.e., how often did you do . . .) and an *effectiveness* scale (i.e., how well can you do . . .). Factor (principal component) analysis of these survey items identified three dimensions: critical thinking, communication, and leadership/teamwork. The three scale scores derived from the principal components produced noteworthy differences (i.e., effect sizes) between seniors and freshmen. The frequency scale produced effect sizes that ranged from 0.21 for mathematics to 0.38 for Teamwork. The effectiveness scale's effect sizes ranged from 0.25 for mathematics to 0.45 for communication. These differences between freshman and senior

responses are in the same substantial change ranges observed in the Test Validity Study portion of this grant for instruments measuring learning outcomes.

Survey items focused on workplace skills demonstrated an ability to readily distinguish between freshman and senior skills. Institution-level analyses revealed that there were noteworthy differences in senior means for both frequency and effectiveness items. In addition these survey items generated difference (senior versus. freshman) scores for institutions that varied substantially across the institutions in the pilot study. This is an important element of analysis that will allow institutions to begin questioning what they are doing that yields results above or below a comparative norm. These findings suggest that the scales can discriminate among institutions in terms of absolute levels of performance (for seniors) and differences between seniors and freshmen.

With respect to civic-engagement survey items, the principal components analysis indicated that it was possible to construct three highly reliable scales: frequency of civic engagement, effectiveness of civic engagement, and beliefs about community. Unfortunately, the statistical loadings for items on the beliefs-about-community scale were so high as to suggest the items didn't measure different aspects of the construct. The use of the dual frequency/effectiveness scales for civic engagement was highly reliable, but taken together they were not able to clearly differentiate among freshmen and seniors. Of the two scales, the effectiveness scale holds the most promise for evaluating differences between seniors and freshmen. A problematic finding in the analysis of civic-engagement items indicates that the frequency of civic engagement declines markedly for seniors.

The questions on citizenship behaviors appear to hold great promise. There is a clear progression from informing, to discussing, to promoting, to working. Differences are most

pronounced for local and state elections. (This finding led to speculation about the contemporaneous impact of the presidential elections that had stimulated so much interest among young voters.). Predictably, voting in civic elections is much higher for seniors because of age and eligibility factors. The exception to these patterns is student government, where freshmen reported much higher levels of participation.

Within the domain of global skills, questions about uses and reliance on diverse sources of information show greater sophistication for seniors. The data show that seniors are more likely to rely on non-U.S. news sources and are less likely to rely on family, friends, politicians, etc. Seniors also report they are more likely to rely on experts (e.g., teachers, scientists, other experts).

Survey items concerned with foreign-language proficiency may be very useful in reporting absolute levels of competence for undergraduate students. For all three kinds of proficiency questions (e.g., reading, writing, and conversational abilities) there is a clear progression of difficulty levels. The exception to this pattern was for items asking if the student has "native or near-native ability." These particular items did not correlate well with the other difficulty-progression level items.

The use of open-ended questions in the survey provoked a considerable amount of debate among the instrument developers. There were questions about whether students would be willing to fill them out, whether the results would be subject to effective analytical treatment, and where such items should reside in the survey instrument. Intense discussion of these issues was largely resolved by two factors. First, the survey designers wanted to include open-ended questions to allow institutional leaders to cull memorable stories about student experience for use in external communications but did not anticipate any substantive statistical treatment of the content of the responses. Second, students

participating in the focus-group reviews of early drafts of the survey made it immediately clear that the open-ended questions should be adjacent to items asking for quantitative information about work and civic-engagement experiences. The field tests show that the open-ended questions on work experiences and civic engagement appear to provide a great deal of information that institutions will want to report. In terms of functionality, it is clear that students do answer the open-ended questions and that they are eager to express what they gained from the work and civic-engagement experiences.

Future Plans

While much has been achieved in developing a survey instrument capable of comprehensively and systematically collecting information about undergraduate learning outcomes that impact the public good, it is equally clear that the survey is in need of some further developmental work. Some scale problems noted above need to be worked out, the questions on beliefs about community need to be redone to gather information about different aspects in that domain, and some survey items that yield repetitive information can be dropped to shorten the survey.

The next steps for *Degrees of Preparation* are further modification of the survey in light of the field testing that has been completed and finding a permanent home for the survey so that it can be developed further, marketed, and made useful throughout the higher-education community. Plans are already under way to address those issues.

Valid Assessment of Learning in Undergraduate Education

Terrel L. Rhodes

The VALUE Project Overview

Valid Assessment of Learning in Undergraduate Education (VALUE) focused on the national conversation around student-learning outcomes and the quality of achievement across a set of important learning outcomes. As part of the Association of American Colleges and University's Liberal Education and America's Promise (LEAP) initiative, the VALUE project built on a philosophy of learning assessment that privileges multiple expert judgments of the quality of student work over reliance on standardized tests administered to samples of students outside of their required curriculum. The project was an effort to focus the national conversation about student learning on the set of essential learning outcomes that faculty, employers, and community leaders say are critical for personal, social, career, and professional success in this century and this global environment. The assessment approaches that VALUE advanced are based on the shared understanding of faculty and academic professionals on campuses across the country.

VALUE assumes that:

■ to achieve a high-quality education for all students, valid assessment data are needed to guide planning, teaching, and improvement. This means that the work students do in their courses and the co-curriculum is the best authentic representation of their learning;

Terrel L. Rhodes is Vice President for the Office of Quality, Curriculum and Assessment at the Association of American Colleges and Universities.

- colleges and universities seek to foster and assess numerous essential learning outcomes beyond the three or four addressed by currently available national standardized tests;

- learning develops over time, is non-linear and should become more complex and sophisticated as students move through their curricular and co-curricular educational pathways within and among institutions toward a degree or similar credential;

- good practice in assessment requires multiple assessments over time; well-planned electronic portfolios provide opportunities to collect data from multiple assessments across a broad range of learning outcomes and modes for expressing learning, while guiding student learning and building reflective self-assessment capabilities;

- assessment of the student work in e-portfolios can inform programs and institutions on their progress in achieving expected goals for external reporting and, at the same time, provide faculty with information necessary to improve courses and pedagogy.

Project Activities

VALUE's work was guided by a national advisory board that was comprised of recognized researchers and campus leaders knowledgeable about the research and evidence on student achievement of key learning outcomes and best practices currently used on campuses to achieve and measure student progress. VALUE focused on the development of rubrics for 15 of the essential learning outcomes that articulate the shared expectations for student performance, derived from faculty and employers across the country. Evidence for the achievement and assessment of these outcomes is demonstrated in the context of the required college curriculum (and co-curriculum), and included models for e-portfolios and rubrics describing ascending levels of accomplishment (beginning, intermediate, and advanced).

VALUE Leadership Campuses

VALUE initially selected 12 leadership campuses to participate in the project based on established use of student e-portfolios on their campuses to assess student learning. Campuses were selected because they used e-portfolios in different ways and in different places in the curriculum. Each VALUE leadership campus used e-portfolio systems in which students collect coursework and related activities in their curricular and co-curricular lives. Upon acceptance into the project, the leadership campuses agreed to test the rubrics developed through VALUE on student e-portfolios on their respective campuses and to determine the usefulness of the rubrics in assessing student learning across the breadth of essential outcomes. In addition, each leadership campus agreed to provide faculty feedback on the usefulness, problems, and advantages of each rubric they tested.

VALUE Partner Campuses

As the rubric-development process proceeded and leadership campuses tested the rubrics, other campuses became aware of the project and began requesting permission to use the rubrics on their campuses. While many of these campuses did not use e-portfolios, they did have collections of student work on which they wished to test the rubrics and provide the project with feedback. As a result of sharing rubrics with this second set of institutions, faculty and others on 100 different campuses tested one or more VALUE rubrics with their students' work.

Learning Outcomes for the Development of Institutional or Meta Rubrics

The essential learning outcomes[2] addressed in the project and for which rubrics were developed fell into three areas:

Intellectual and Practical Skills:
- Inquiry and analysis
- Critical thinking

- Creative thinking
- Written communication
- Oral communication
- Quantitative literacy
- Information literacy
- Teamwork
- Problem solving
- Reading

Personal and Social Responsibility:
- Civic knowledge and engagement—local and global
- Intercultural knowledge and competence
- Ethical reasoning
- Foundations and skills for lifelong learning

Integrative Learning:
- Integrative learning

Process of Developing Rubrics

As part of the VALUE project, teams of faculty, academic professionals, and assessment experts gathered, analyzed, synthesized, and drafted rubrics based on a collection of existing campus rubrics and related materials for the 15 outcomes, to create what we initially called meta rubrics, or shared expectations for learning. The meta rubrics are simply statements of key criteria or characteristics of the particular learning outcome; statements of what demonstrated performance for each criterion looks like at four levels are displayed in a one-page table (see example below). The VALUE rubrics are "meta" in the sense that they synthesize the common criteria and performance levels gleaned from numerous individual campus rubrics and are synthesized into general rubric tables for each essential learning outcome. Each meta rubric contains the key criteria most often found in the many campus rubrics collected, and represents a carefully considered summary of criteria widely considered critical to judging the quality of student work in each outcome area.

The rubric-development process was a proof of concept. The claim was that faculty and other academic and student-personnel professionals do have fundamental, commonly held expectations for student learning, regardless of type of institution, disciplinary background, part of the country, or whether the college is public or private. Further, these commonly shared expectations for learning can also be articulated for developmentally more-challenging levels of performance and demonstration.

The process of reviewing collections of existing rubrics, joined with faculty expertise across the range of outcomes, uncovered the extent to which there were similarities among campuses on core learning expectations. By identifying outcomes in terms of expectations for demonstrated student learning among disparate campuses, a valuable basis for comparing levels of learning through the curriculum and co-curriculum emerged. This is especially useful as students, parents, employers, and policy makers seek valid representations of student academic accomplishments, especially when the expected learning can be accompanied by examples of actual student work that tangibly demonstrates learning.

The rubric teams began developing draft meta rubrics in spring 2008. By late spring, three rubrics had been drafted. Those three rubrics were then pilot tested by faculty on some of the leadership campuses. Feedback from the first round of testing was used by the respective teams to engage in a second round of drafting and redrafting the rubrics. By fall 2008, drafts of the rubrics articulating 14 essential learning outcomes were in place. In early 2009, the new rubric drafts were piloted on both leadership and partner campuses across the country. Also, a 15th rubric, on reading, was developed in spring 2009 at the request of rubric-development team members and campus faculty. In late spring 2009, the rubrics underwent another round of campus testing. A final "tweaking" of the rubrics occurred in summer 2009. In

September 2009, the VALUE rubrics were released for general use. (see aacu.org/value/rubrics/index.cfm).

E-portfolios as the Mode for Presenting Students' Work

E-portfolios were chosen as the medium for collecting and displaying students' work for three primary reasons: (1) there were sufficient numbers of campuses using e-portfolios for assessment of learning to represent multiple sectors and types of institutions; (2) it was easier to share student work among campuses, faculty teams, and evaluators digitally than to transport groups of people; and (3) e-portfolios allowed learning to be presented using a broad range of media to capture the multiple ways in which we learn and can demonstrate our learning. E-portfolios provided both a transparent and portable medium for showcasing the broad range of complex ways students demonstrate their knowledge and abilities for purposes such as graduate school and job applications, as well as to benchmark achievement among peer institutions. To ensure that judgments about student learning reflect the learning that actually occurs on our campuses, the student artifacts were drawn primarily from the work students complete through their required curriculum and co-curriculum.

The e-portfolio is an ideal format for collecting evidence of student learning, especially for those outcomes not amenable to or appropriate for standardized measurement. Additionally, e-portfolios can facilitate students' reflection upon and engagement with learning across multiyear degree programs, across different institutions, and across diverse learning styles, while helping students to set and achieve personal learning goals.

The rubric development teams endeavored to craft language in the rubrics that would not be text bound, but open to use for learning performances that were graphical, oral, video, digital, etc. VALUE rubrics attempt to reflect the research and the reality of today's students and the learning environments that engage us all in

technological, social, and extra-campus learning that is integral to the learning that occurs in the more traditional, formal classroom.

A Final Piece of the Project

Since it was important that the rubrics and the e-portfolio collections of student work serve both campus assessment and non-campus accountability purposes, VALUE engaged a national panel to review the rubrics, use the rubrics to assess student e-portfolios, and provide feedback on the usefulness of the rubrics and the student e-portfolios. The national panel included faculty and administrators who were familiar with rubrics and e-portfolios, but who were not involved in the VALUE project; faculty and administrators who were familiar with neither rubrics nor e-portfolio usage; and selected employers, policy makers, parents, teachers, and community leaders.

The panel used three rubrics (one from each category of the learning outcomes, specifically critical thinking, ethical reasoning and integrative learning) to assess the same set of student e-portfolios. The e-portfolios represented students' work from different types and sizes of institutions, different majors, and different years in school. The panel engaged in a process establishing inter-rater reliability. The panels found two of the rubrics to be usable and useful in assessing student work. A third was found to be usable but in need of revision and clarification of language. There was a high degree of agreement on the performance levels of the students. The panel found that the rubrics represented important dimensions of learning. The results of their reviews and their feedback was used by the rubric-development teams for the final "tweaking" of the rubrics. The national panel was an initial indicator of the rubrics' ability to communicate similar meaning about quality of learning to very differently positioned sets of people both inside and outside the academy.

Conclusion

The VALUE rubrics are meant to both capture the foundations of a nationally shared set of meanings around student learning, and to be useful at both general institutional and programmatic levels. The VALUE rubrics, as written, must be translated by individual campuses into the language, context, and mission of their institutions. Programs and majors will have to translate the rubrics into the conceptual and academic constructs of their particular area or discipline. Individual faculty will have to translate the rubrics into the meaning of their assignments and course materials in order for the rubrics to be used effectively to assess their student assignments.

However, as institutional versions of the rubrics are mapped onto the VALUE rubric criteria and performance levels, each level of the institution—individual faculty, disciplines, programs, and institution-wide—now can have confidence that their assessments are not idiosyncratic, but rather exist within a national understanding of learning expectations and demonstrated quality. This translation to the local parlance allows for the work of students and faculty on specific assignments in specific courses to not only serve the purposes of assigning grades and performance indicators in a course, but also for the same pieces of work and their assessment to be sampled and/or aggregated for program-review or assessment purposes, and ultimately at an institutional level. Through this deconstruction and construction process, the rubrics become useful to faculty and students on the ground on a day-to-day basis for moving through a course of study. Through aggregating and sampling, the exact same work can also be used to provide a macro review of student learning without having to start anew or devise separate modes of gathering assessment data. Multiple purposes and needs can be met through shared, layered, and textured rubrics, facilitating both formative assessments for learning and summative assessment for accountability reporting.

Plans are under way to work with campuses and disciplinary associations to develop the VALUE rubrics for use within a set of major programs, reflecting the concepts, language, and content of the disciplines, while maintaining the core criteria for learning developed through the meta rubric process. Washington State University has begun this process for critical thinking. In addition, several e-portfolio companies already have adopted the VALUE rubrics as organizing frameworks for their e-portfolio products and are finding that many of their user-campuses are employing the rubrics for advancing and assessing student learning.

As stated earlier, VALUE is a first step, a proof of concept. The evidence supports the finding that faculty and institutions can talk about a broadly shared understanding of learning across a broad range of outcomes and at increasingly more challenging levels of performance. We are learning that assessment of student learning can be rigorous, effective, useful, and efficient. There is integrity and face and use validity in the meta rubrics and portfolio assessment that can lead to rich evidence of student learning to meet demands for accountability, and at the same time encourage improvements in teaching and learning for faculty and staff. Perhaps most important, this process can allow students to develop their own abilities to engage in self-assessment and meaning making.

Acknowledgments

INTEGRATIVE LEARNING VALUE RUBRIC

for more information, please contact value@aacu.org

A|A C&U Association of American Colleges and Universities
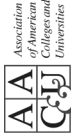
Definition

Integrative learning is an understanding and a disposition that a student builds across the curriculum and co-curriculum, from making simple connections among ideas and experiences to synthesizing and transferring learning to new, complex situations within and beyond the campus.

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone 4	Milestones 3	Milestones 2	Benchmark 1
Connections to experience *Connects relevant experience and academic knowledge*	Meaningfully **synthesizes** connections among experiences outside of the formal classroom (including life experiences and academic experiences such as internships and travel abroad) to **deepen understanding** of fields of study and to broaden own points of view.	Effectively **selects and develops** examples of life experiences, drawn from a variety of contexts (e.g. family life, artistic participation, civic involvement, work experience), to **illuminate** concepts/theories/frameworks of fields of study.	**Compares** life experiences and academic knowledge to infer differences, as well as similarities, and **acknowledge perspectives** other than own.	**Identifies** connections between life experiences and those academic texts and ideas **perceived as similar and related** to own interests.
Connections to discipline *Sees (makes) connections across disciplines, perspectives*	Independently creates wholes out of multiple parts (synthesizes) or draws conclusions by combining examples, facts, or theories from more than one field of study or perspective.	Independently connects examples, facts, or theories from more than one field of study or perspective.	When prompted, connects examples, facts, or theories from more than one field of study or perspective.	When prompted, presents examples, facts, or theories from more than one field of study or perspective.
Transfer *Adapts and applies skills, abilities, theories, or methodologies gained in one situation to new situations*	When prompted, presents examples, facts, or theories from more than one field of study or perspective.	Adapts and applies skills, abilities, theories, or methodologies gained in one situation to new situations **to solve problems or explore issues.**	Uses skills, abilities, theories, or methodologies gained in one situation in a new situation **to contribute to understanding of problems or issues.**	Uses, in a basic way, skills, abilities, theories, or methodologies gained in one situation **in a new situation.**
Integrated Communication	Fulfills the assignment(s) by choosing a format, language or graph (or other visual representation) **in ways that enhance meaning,** making clear the interdependence of language and meaning, thought and expression.	Fulfills the assignment(s) by choosing a format, language or graph (or other visual representation) **to explicitly connect content and form,** demonstrating awareness of purpose and audience.	Fulfills the assignment(s) by choosing a format, language or graph (or other visual representation) that **connects in a basic way** what is being communicated (content) with how it is said (form).	Fulfills the assignment(s) (i.e. to produce an essay, a poster, a video, a powerpoint presentation, etc.) **in an appropriate form.**
Reflection and Self Assessment *Demonstrates a developing sense of self as a learner, building on prior experiences to respond to new and challenging contexts (may be evident in self assessment, reflective, or creative work)*	Envisions a future self (and possibly makes plans that build on past experiences) that have occurred across multiple and diverse contexts.	Evaluates changes in own learning over time, recognizing complex contextual factors (e.g. works with ambiguity and risk, deals with frustration, considers ethical frameworks).	Articulates strengths and challenges (within specific performances or events) to increase effectiveness in different contexts (through increased self awareness).	Describes own performances with general descriptors of success and failure.

Interpretation of Findings of the Test Validity Study Conducted for the Voluntary System of Accountability

David Shulenberger and Christine Keller

Introduction

This document reflects the perspective of leaders of the Voluntary System of Accountability's (VSA) on how the findings of the test validity study (TVS) inform measurement of learning outcomes within the VSA. The TVS report presents findings at both the institutional level and the individual student level; this abstract focuses primarily on institutional-level findings except where reference to student-level findings is necessary to fully understand institutional-level results. It is not intended to be a summary of the TVS report. That report and its executive summary can be found on the VSA Web site at voluntarysystem.org/index. cfm?page=research.

In this abstract, four questions of potential concern to VSA participants are posed and relevant findings of the TVS report are reported under each question. (The TVS research questions[3] are broader than the ones we examine here and are listed in the footnote). Excerpts from the TVS report are generally quoted verbatim and are printed in different type, with the page reference following in parentheses so that the reader can easily find the material quoted in the body of the TVS report. We have used boldfaced type for phrases from the TVS report that most directly bear on the question under discussion. We stress that the

---

David E. Shulenburger is Vice President, Academic Affairs and Christine M. Keller, is Director of Research Policy and Analysis and Executive Director, Voluntary System of Accountability at APLU.

interpretations of TVS findings contained in this document are those of the authors.

Background

Two taskforces[4] of higher-education leaders from a variety of backgrounds thoroughly evaluated 16 potential tests of learning outcomes and recommended three for use of institutions participating in the Voluntary System of Accountability. The VSA presidential advisory board carefully reviewed and ultimately confirmed the taskforce's recommendation. Multiple test options were identified for use in VSA because public universities expressed strong desire to have the ability to select a test best suited to their particular campus circumstances. The three tests chosen were:

■ Collegiate Assessment of Academic Proficiency (CAAP)—two modules: critical thinking and writing an essay. CAAP is a product of ACT.

■ Collegiate Learning Assessment (CLA)—the complete test, including a performance task and an analytic writing task (consisting of a make-an-argument and a critique-an-argument prompt). The CLA measures critical thinking, analytic reasoning, problem-solving, and written communication. CLA is a product of the Council for Aid to Education (CAE).

■ Measure of Academic Proficiency and Progress (MAPP)— two subscores of the test: critical thinking and written communication. MAPP is a product of the Educational Testing Service (ETS).

The taskforces determined that the CAAP, CLA, and MAPP were valid tests for measurement of critical thinking and written communication.[5] Two types of validity need to be distinguished: face validity and construct validity. The VSA taskforce concluded that the portions of the three tests selected for use in VSA had face validity. In other words, each of the tests presents the test taker with tasks that clearly require the use of critical thinking

and written communication abilities. Face validity is very important as those considering the results must be confident that the skills being measured are those relevant to and valued by future employers. However, the VSA taskforces recommended additional research to evaluate the concurrent validity across the three tests so the VSA could more confidently state that learning-outcomes results were generally comparable for each of the different test options.

In fall 2007, the Fund for the Improvement of Postsecondary Education (FIPSE) funded a test validity study of the three tests of critical thinking and written communication used to measure value-added student learning outcomes within the VSA. The Association of Public and Land-grant Universities (APLU), under the direction of FIPSE grant co-principal investigator David Shulenburger, subcontracted for the testing and analytical work to be done by a consortium of testing experts led by Stephen Klein from the Council for Aid to Education (CAE), Ou Lydia Liu from Educational Testing Service (ETS), and James Sconing from ACT.

Test Frame

In fall 2008 and spring 2009, 13 tests were administered to approximately 1,100 students at 13[6] colleges and universities across the U.S.[7] The tests included the portions of CAAP, MAPP, and CLA used in VSA,[8] along with additional component tests of CAAP and MAPP: two tests in reading, two tests in mathematics and one in science. The tests and constructs are outlined in Table 1, reproduced from the full TVS report. *(Klein, Liu, Sconing, Bolus, Bridgeman, Kugelmass, Nemeth, Robbins, and Steedle, 2009)*

Table 1.
Summary of Constructs and Corresponding Tests

Construct(s)	Tests
Critical Thinking	MAPP Critical Thinking, CAAP Critical Thinking, CLA PT*, CLA CA*
Writing	MAPP Writing, CAAP Writing Skills, CAAP WritingEssay*, CLA MA*
Mathematics	MAPP Mathematics, CAAP Mathematics
Reading	MAPP Reading, CAAP Reading
Science	CAAP Science

*Indicates constructed-response test format.

Each of the 13 institutions recruited a sample of 46 first-time, full-time freshmen and 46 seniors who had entered the institution as freshmen to take the test.[9] Student participants were given a $150 Amazon.com gift certificate if they completed three separate testing sessions.

Key Findings for the VSA

1. What is the reliability of school-level scores of different measures of writing and critical-thinking ability?

Overall, the reliability of school-level scores was high across all the measures of writing and critical-thinking abilities. The TVS report explains the implications of high reliability at the school level in two different sections:

> School-level reliability refers to score consistency (i.e., a school receiving a similar mean score regardless of the sample of students taking the test). Reliability is reported on a scale from 0.00 to 1.00, where higher values indicate greater reliability. **With schools as the unit of analysis, score reliability was high on all 13 tests (mean was 0.87 and the lowest value was 0.75). Thus, score reliability is not a major concern when using school level results with sample sizes comparable to those obtained for this study.** *(Klein, et al., 2009, p. 4)*

> The school-level reliability coefficients indicate that scores from these tests are adequately reliable by most standards. A few coefficients are smaller than would typically be observed, but these anomalous values may simply reflect instability of estimates in the small sample of colleges. Generally, the school-level reliabilities were high (greater than 0.90), and this bodes fairly well for the use of relatively small samples for institutional assessment. The within-school sample sizes never exceeded 50 students for MAPP and never exceeded 30 for CLA or CAAP. It

should be noted, however, that the between-school variance was quite large given the small number of schools, which would have a positive impact on school-level reliability. *(Klein, et al., 2009, pp.28-29)*

Table 5 *(Klein, et al., 2009, p. 29)* from the TVS report details the specific reliability coefficients and is reproduced below.

The TVS report's observation regarding sample size requires clarification for institutions participating in the VSA. As part of the VSA guidelines for administering one of the learning-outcomes tests, participants are instructed to follow the recommendations of the appropriate testing company. At a minimum, CLA users are

Table 5.
School–level reliabilities computed as the mean of 1,0 random Spearman–Brown adjusted split–half reliabi

Measure	Freshman	Senior
MAPP Critical Thinking	0.95	0.91
CAAP Critical Thinking	0.86	0.88
CLA Performance Task	0.85	0.64
CLA Critique–an–Argument	0.86	0.84
MAPP Writing	0.94	0.88
CLA Make–an–Argument	0.87	0.81
CAAP Writing Skills	0.92	0.84
CAAP Writing Essay	0.68	0.82
MAPP Mathematics	0.95	0.93
CAAP Mathematics	0.93	0.90
MAPP Reading	0.94	0.88
CAAP Reading	0.92	0.83
CAAP Science	0.92	0.92

advised to test a minimum sample size of 100 each for freshmen and seniors; MAPP and CAAP users are advised to test a minimum of 200 each for freshmen and seniors. All three test companies recommend larger samples when a school wants to disaggregate the results by student groups. Thus the high correlations and reliable results obtained in the TVS study with samples of 30 to 50 students are useful for purposes of validation, but VSA universities should continue to follow established minimums of 100 or 200 for their value-added measurement.

2. **To what degree do different measures designed to assess the same construct (such as critical thinking) correlate with each other as compared to tests that are designed to assess other constructs (such as reading)?**

It would be exceedingly unusual to find a test that measures only a single, unique ability. For example, essay writing and critical-thinking skills are clearly intertwined. Science and math tests draw on critical-thinking skills as do tests of reading comprehension. Math "word-problems" require a certain level of reading comprehension as well as mathematical skills. As the researchers in the TVS study state "it is recognized that a single test may measure multiple constructs and that constructs may overlap." *(Klein, et al., 2009, p. 11)* In addition, individuals who are proficient in one domain may be proficient in another domain. For these reasons test scores generally exhibit a significant level of covariance (i.e., the test scores move in tandem). The TVS researchers describe the complexity of interpreting correlations among constructs in the following excerpt.

> This portion of the TVS sought evidence of convergent and discriminant validity. Evidence of convergent validity is obtained when a test has high correlations with other measures of the same (or a similar) construct. Evidence of discriminant validity is obtained when a test has lower correlations with measures of different constructs than it has with tests assessing the same construct. Such evidence helps confirm that tests measuring the same construct should be highly correlated, but a high correlation between two tests does not mean that they measure the same construct. It means only that students with the skills required to perform well on one test tend to have the skills required to perform well on the other test. *(Klein, et al., 2009, p. 20)*

The basic correlation matrices in the TVS tables 2a and 2b *(Klein, et al., 2009, p. 24)* are reproduced below. Both the student- and

school-level results are shown because the student-level data inform the conclusions concerning the school-level results.

As demonstrated in Table 2a, the correlation patterns for the student-level results generally supported the construct validity among the different measures. As detailed in the TVS report:

> On the whole, patterns of student-level correlations revealed that the TVS measures correlated most highly with measures of similar constructs (e.g., critical thinking correlating with critical thinking, writing with writing, reading with reading, and math with math). *(Klein, et al., 2009, p. 24)*

> **. . . results were consistent with the conclusion that tests purporting to measure the same or similar constructs do indeed measure those constructs (and not other constructs).** Specifically, an examination of the student-level correlations revealed that two tests of the same construct

Table 2a.
Student–level correlation matrix with standard correlations shown above the diagonal

Construct(s)	Test	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
Critical Thinking	1. MAPP		0.75	0.53	0.52	0.76	0.45	0.68	0.34	0.63	0.46	0.86	0.76	0.74
	2. CAAP			0.58	0.47	0.66	0.39	–	0.32	0.57	–	0.71	–	0.74
	3. CLA PT				–	0.50	–	0.49	0.32	0.46	0.40	0.55	0.52	0.52
	4. CLA CA					0.48	0.47	0.49	0.40	0.46	0.44	0.49	0.50	0.50
Writing	5. MAPP						0.44	0.72	0.33	0.60	0.51	0.73	0.70	0.63
	6. CLA MA							0.44	0.37	0.40	0.39	0.43	0.46	0.39
	7. CAAP								–	0.58	0.48	0.70	0.71	–
	8. CAAP Ess.									0.29	–	0.31	–	0.28
Mathematics	9. MAPP										0.76	0.60	0.55	0.71
	10. CAAP											0.46	0.44	–
Reading	11. MAPP												0.76	0.70
	12. CAAP													–
Science	13. CAAP													

Table 2b.
School–level correlation matrix with standard correlations shown above the diagonal and reliabilities shown on the diagonal

Construct(s)	Test	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
Critical Thinking	1. MAPP	0.93	0.93	0.83	0.93	0.96	0.85	0.89	0.62	0.95	0.93	0.96	0.82	0.93
	2. CAAP		0.87	0.79	0.87	0.94	0.79	0.91	0.75	0.90	0.86	0.93	0.76	0.95
	3. CLA PT			0.75	0.73	0.84	0.67	0.77	0.58	0.91	0.91	0.90	0.76	0.86
	4. CLA CA				0.85	0.92	0.90	0.90	0.61	0.82	0.77	0.91	0.91	0.79
Writing	5. MAPP					0.91	0.86	0.97	0.70	0.92	0.90	0.96	0.87	0.90
	6. CLA MA						0.84	0.83	0.67	0.74	0.72	0.82	0.86	0.69
	7. CAAP							0.88	0.74	0.83	0.78	0.93	0.89	0.81
	8. CAAP Ess.								0.75	0.57	0.56	0.62	0.71	0.61
Mathematics	9. MAPP									0.94	0.98	0.94	0.71	0.98
	10. CAAP										0.92	0.91	0.70	0.96
Reading	11. MAPP											0.91	0.86	0.91
	12. CAAP												0.88	0.65
Science	13. CAAP													0.92

usually correlated higher with each other than they did with measures of other constructs provided the response format was taken into consideration. *(Klein, et al., 2009, p.30)*

The correlation patterns at the school level (Table 2a) were parallel to the patterns at the student level but less distinct. The TVS report explains this finding in more detail.

The pattern of results at the school level was much fainter because all the correlations were much higher and the differences among them much smaller. This came about as a result of the much higher level of score reliability for all the measures at the school level. *(Klein, et al., 2009, p. 31)*

For example, the mean correlation between two multiple-choice tests of the same construct (r = .94) at the school level was only very slightly higher than the mean correlation between two multiple-choice tests of different constructs (r = .92). *(Klein, et al., 2009, p. 31)*

The mean correlation between two constructed-response tests of the same construct (r = .84) at the school level was only slightly higher than the mean correlation between two constructed-response tests of different constructs (r = .83). *(Klein, et al., 2009, p. 31)*

In addition, the mean correlation between multiple-choice and constructed-response tests of critical thinking (r = .89) was only slightly higher than it was between constructed-response and multiple-choice tests of different constructs (r = .85) or among constructed-response tests of different constructs (r = .83). There also continued to be a lower correlation between multiple-choice and constructed-response tests of writing (r = .83). *(Klein, et al., 2009, p. 31)*

Thus, while there was less differentiation among the coefficients, the pattern of results at the school level was consistent with the pattern at the student level. *(Klein, et al., 2009, p. 31)*

3. Is the average difference in mean scores (effect sizes) between freshmen and seniors similar across the different measures of the same construct?

In order to compare changes in mean scores across tests with dissimilar score distributions and to control for differences in average student ACT or SAT scores, the researchers created a standardized index of "effect size." The effect size reflects the average difference between freshmen and seniors on the TVS tests. Larger effect sizes indicate greater differences in mean scores between freshmen and seniors.

The test validity study found the average difference in mean scores between freshmen and seniors to be nearly identical across different measures of the same construct.

Effect sizes were not systematically related to the constructs tested, response format, or test publisher. For example, **the average effect size across constructs for the ACT, CAE, and ETS measures were 0.33 (excluding mathematics), 0.31, and 0.34, respectively.** *(Klein, et al., 2009, p. 4)*

The TVS analyses include both observed and adjusted effect sizes. An adjustment was necessary because on average seniors had higher ACT or SAT scores than freshmen. Adjusting the effect sizes created a standardized measure that could be interpreted to reflect learning gains during college rather than prior academic achievement.

The observed (or unadjusted) effect size results are described in more detail below and shown in Table 4a. *(Klein, et al., 2009)*

The observed (unadjusted) effect sizes and their corresponding 95 percent confidence intervals provided in Table 4a (and displayed in Figure 1a) indicate that **there were significant differences between the freshmen and seniors on all measures except CAAP Mathematics**. Recall, however, that some component of the positive effect sizes reflects differences in entering ability rather than learning that took place during college. Across the TVS measures (excluding CAAP Mathematics), the average effect size was 0.42, and the average difference in ability between freshmen and seniors (as measured by the SAT or ACT) reflected an effect size of 0.10. This suggests that 24 percent (.10/.42) of the observed freshman-senior difference can be accounted for by entering ability differences. *(Klein, et al., 2009, p. 27)*

The adjusted effect size results are described in the paragraph below.

Adjusted effect sizes, which control for differences in entering ability, are provided in Table 4b and displayed in Figure 1b. The adjustment tends to make the effect sizes smaller and the 95 percent confidence intervals larger. Although three adjusted effect sizes were not significantly different from zero (CLA Performance Task, CAAP Writing Essay, and CAAP Mathematics), all adjusted effect size estimates were positive except for CAAP Mathematics, which indicates that the TVS measures are sensitive to the increase in skills that occurs over the course of college. The largest adjusted effect sizes were 0.46 for MAPP Critical Thinking, 0.46 for CAAP Reading, 0.45 for MAPP Reading, and 0.40 for CLA Critique-an-Argument. Figure 1b shows that the confidence intervals for all positive adjusted effect sizes overlap somewhat, and this suggests that many differences in adjusted effect sizes were not statistically significant. This was especially true of the writing tests, which had adjusted effect sizes ranging from 0.22 to 0.32. The MAPP and CAAP Reading tests also had very similar adjusted effect

Table 4a.
Precision–weighted average observed effect sizes

Measure	d_+	$se(d_+)$	95% Conf. Interval	
			Lower	Upper
MAPP Critical Thinking	0.57	0.064	0.44	0.69
CAAP Critical Thinking	0.48	0.091	0.30	0.65
CLA Performance Task	0.47	0.090	0.30	0.65
CLA Critique–an–Argument	0.39	0.090	0.22	0.57
MAPP Writing	0.34	0.063	0.22	0.46
CLA Make–an–Argument	0.28	0.089	0.10	0.45
CAAP Writing Skills	0.36	0.090	0.18	0.54
CAAP Writing Essay	0.37	0.092	0.19	0.55
MAPP Mathematics	0.32	0.063	0.19	0.44
CAAP Mathematics	–0.12	0.089	–0.29	0.06
MAPP Reading	0.55	0.064	0.42	0.67
CAAP Reading	0.48	0.091	0.31	0.66
CAAP Science	0.49	0.091	0.31	0.67

Figure 1a.
Precision–weighted average observed effect sizes
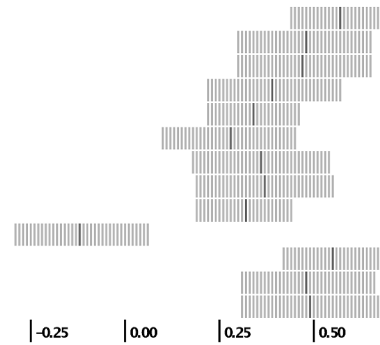
Table 4b.
Precision–weighted average adjusted effect sizes

Measure	$d_{+,adj}$	$se(d_{+,adj})$	95% Conf. Interval	
			Lower	Upper
MAPP Critical Thinking	0.46	0.089	0.29	0.64
CAAP Critical Thinking	0.31	0.128	0.06	0.56
CLA Performance Task	0.23	0.127	–0.02	0.48
CLA Critique–an–Argument	0.40	0.126	0.15	0.65
MAPP Writing	0.24	0.089	0.06	0.41
CLA Make–an–Argument	0.29	0.126	0.04	0.54
CAAP Writing Skills	0.32	0.127	0.07	0.57
CAAP Writing Essay	0.22	0.130	–0.03	0.48
MAPP Mathematics	0.22	0.089	0.04	0.39
CAAP Mathematics	–0.15	0.127	–0.40	0.09
MAPP Reading	0.45	0.089	0.27	0.62
CAAP Reading	0.46	0.129	0.21	0.71
CAAP Science	0.33	0.128	0.08	0.58

Figure 1b.
Precision–weighted average adjusted effect sizes
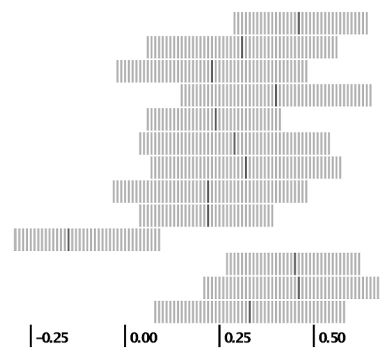
sizes (0.45 and 0.46, respectively). There was greater variation among the tests that measure critical thinking skills. *(Klein, et al., 2009, p. 27)*

4. **Do the scores on tests that use different response modes (such as essay versus multiple choice) to assess a given competency (such as writing ability) correlate more highly with each other than they do with scores on tests that use the same response mode but assess different constructs?In other words, to what extent are the correlations among tests a function of mastery of the constructs being measured and the response modes of the tests?**

The relative consistency in effect size across the three tests provides evidence that differences in score gains are associated with learning differences and not with the test or test format. More specifically:

> Effect sizes ranged from approximately one quarter to one half of a standard deviation. Furthermore, **effect sizes were fairly consistent across tests, test formats (multiple-choice and constructed-response), test publishers (ACT, CAE, and ETS), and constructs.** *(Klein, et al., 2009, p.32)*

Key Points for VSA Participants

The TVS findings provide evidence that across test constructs, response formats, and test publishers:

- Correlations are generally high at the school level
- Adjusted effect sizes are consistent
- School level reliabilities are high

The results suggest that when the analysis is conducted at the school level, all the tests order schools similarly, regardless of which constructs they are designed to measure or which response format is used.

The TVS findings allow leaders at VSA institutions to select the instrument that best fits the circumstances at their particular institution with confidence in the technical and measurement abilities of all three options. Other important considerations are described by the TVS authors.

> Finally, given the findings above and particularly the high correlation among the measures, the decision about which measures to use will probably hinge on their acceptance by students, faculty, administrators, and other policy makers. There also may be trade-offs in costs, ease of administration, and the utility of the different tests for other purposes, such as to support other campus activities and services. Indeed,

the assessment program may include guidance on the interpretation of results and their implications for programs and activities that complement the testing program's goal of improving teaching and learning. For this to be accomplished systematically and systemically, adopters of any test covered in this study should also understand the catalytic roles played by campus leadership, willing faculty, and cultures of evidence. Equally important are the benefits inherent in assessment tools that are reliable (correlate well with other tools), have face validity (represent the type of performance you want students to demonstrate), and that couple summative data with formative diagnostics to improve teaching and learning (Klein, et al., 2009, p. 33).

Cautions for VSA Participants

1. The findings of the TVS study demonstrate that the three tests used within the VSA have highly correlated average scores at the school level. The correlations are more varied and generally lower at the student-level. In particular, scores from brief, open-ended tests are less reliable at the student level.

2. Despite the high correlations among the tests measuring the same construct, especially critical thinking, the study does not "prove" that the tests measure the same thing. What the study shows is that students who do well on one test of "critical thinking" generally do well on another test of "critical thinking."

3. Although on average, the tests provide similar adjusted effect sizes (which could be considered a measure of value-added, the TVS did not have adequate data to directly evaluate the comparability of value-added scores. The appropriate conclusion is that each of the three tests provides similar results for ordering schools by their mean test scores.

Sources

Klein, S., Liu, O.L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., Nemeth, A., Robbins, S., and Steedle, J. (September 29, 2009). *Test Validity Study (TVS) Report.* Supported by the Fund for Improvement of Postsecondary Education (FIPSE). Online at voluntarysystem.org/index.cfm?page=research.

Endnotes

[1]Cognitive interview were conducted with a student taking the survey on a Web-connected computer with a validation team member sitting with her or him. The student was asked to say aloud whatever he or she was thinking about the survey wording, mechanics, instructions, and so on. The validation team member also observed what the student was and was not doing (e.g., not reading instructions) on each page.

[2]See College Learning for the New Global Century. 2007. Washington, D.C.: Association of American Colleges and Universities, p.12.

[3]The research questions in the TVS study are: 1. What are the relationships among scores on commonly used college-level tests of general educational outcomes? Are these relationships a function of the specific skills the tests presumably measure, the tests' formats (multiple-choice or constructed-response), or the tests' publishers? 2. Is the difference in average scores between freshmen and seniors related to the construct tested, response format, or the test's publisher? 3. What are the reliabilities of school-level scores on different tests of college learning?

[4]For a full description of the committee process and membership see voluntarysystem.org/index.cfm?page=background and for the full report of the committee see voluntarysystem.org/docs/cp/LearningOutcomesInfo.pdf.

[5]Analytic reasoning is sometimes listed as a third core skill, but there is disagreement as to whether this ability is actually integral to the other two core skills so this document simply refers to two core skills.

[6]The 13 universities and colleges are Alabama A& M University, Arizona State University at the Tempe Campus, Boise State University, California State University, Northridge, Florida State University, Trinity College, Massachusetts Institute of Technology, University of Colorado at Denver, University of Michigan-Ann Arbor, University of Minnesota-Twin Cities, University of Texas at El Paso, University of Vermont, University of Wisconsin-Stout.

[7]The results of the test administration at each university are confidential, and the results will not be presented in any way that serves to identify a specific university's results.

[8]The MAPP writing essay test that is a component of VSA was not administered to students because of the great similarity of it with the CAAP writing essay. This economy was needed in order to enable the full array of tests of different constructs to be included.

[9]1,051 students took all three tests, 23 took only two tests and 51 took only one test. 51 percent of the students taking all three tests were freshmen and 49 percent were seniors, a near perfect distribution. The resulting samples were reasonable reflections of their school's populations. Appendix C of the TVS Report has a full description of the sample and school characteristics.

Appendix A

Degrees of Preparation Workflow

Establishing the Validity for Degrees of Preparation

Three Teams
• Subject Expert Panel (5)
• Validation Team (9)
• Institutional Research Analysis Team (4)

Nov /Dec '07: Subject Experts Fine-tune Topic Areas & Craft Items

Dec '07: Validation Team Designs Process & Reviews Draft Items

Dec '07 & Jan '08: Validation Team Members are paired to create items that address the topic areas & research questions as outlined in the proposal

Feb '08: Focus Groups Protocol is designed and disseminated to Validation Team Members IRB Approval is sought at all team institutions

Jan/Feb '08: Survey items are circulated via email and fine-tuned.

Feb to April '08: Focus Groups are conducted in multiple locations

April '08: Validation Team Members meet to discuss results of student focus groups and further tweak the survey instrument

April '08: Survey instructions and items are polished and the instrument under goes more focus groups

April '08: Results of the last group of focus groups are incorporated

May/June '08: An on-line version is created

May '08: Cognitive interviewing protocol is created and disseminated

June '08: Cognitive interviews (CI) across multiple institutions occur

July '08: Validation Team and Expert Panel meet in DC to discuss and fine-tune the final instrument using the results of the CIs

August '08: Final Survey is created

Obtain IRB Approval

Oct' 08: Field Testing 12 State Institutions

Dec' 08: Analyzed Data & Reviewed findings & fine-tuned survey items

Spring '09: Field Testing Part 2

APPENDIX B

Rising to the Challenge: Meaningful Assessment of Student Learning Project Participants

AAC&U-led VALUE Rubrics Portfolio Evaluation Project

National Advisory Board
- Randy Bass, Assistant Provost for Teaching and Learning Initiatives, Georgetown University, Washington, D.C.
- Marcia Baxter Magolda, Distinguished Professor of Educational Leadership, Miami University, Ohio
- Veronica Boix Mansilla, Research Associate and Lecturer on Education, Harvard University, Massachusetts
- Johnnella Butler, Provost, Spelman College, Georgia
- Helen Chen, Research Scientist, Stanford University, California
- Ariane Hoy, Senior Program Officer, The Bonner Foundation, New Jersey
- George Kuh, Chancellor's Professor and Director, Center for Postsecondary Research, Indiana University Bloomington
- Peggy Maki, Education Consultant; Peggy Maki Associates
- Marcia Mentkowski, Director, Educational Research and Evaluation, Alverno College, Wisconsin
- Gloria Rogers, Associate Executive Director, Professional Services, ABET, Inc., Maryland
- Carol Geary Schneider, President, Association of American Colleges and Universities, Washington, D.C.
- Robert Sternberg, Dean of Arts and Sciences, Tufts University, Massachusetts
- Kathleen Blake Yancey, Kellogg H. Hunt Professor of English, Florida State University

Reading Rubric

- Susan Albertine, Senior Director for LEAP State Initiatives, Association of American Colleges and Universities, Washington, D.C.
- Maureen Erickson, Director of Assessment, Cayuga Community College, New York
- Alan Grose, Administrative Coordinator, Core Seminar Program, Long Island University-Brooklyn Campus, New York
- Sharon Klein, Professor of Linguistics and Director of Writing and Reading Across Disciplines, California State University, Northridge
- P. Pearson, Dean and Professor, Graduate School of Education, University of California, Berkeley

Oral Communication Rubric

- Terry Underwood, Professor of Language and Literacy, Faculty Assessment Coordinator, California State University-Sacramento
- Jo Beld, Director of Evaluation and Assessment and Professor of Political Science, St. Olaf College, Minnesota

Integrative Learning Rubric

- Mary Gill, Associate Dean of the Faculty, Buena Vista University
- Laura Blake, CIRP Assistant Director, Higher Education, Iowa Research Institute, University of California, Los Angeles
- Brad Mello, Associate Director for Educational Initiatives, National Communication Association, Washington, D.C.
- Mark Braun, Senior Vice President for Academic Affairs and Dean of the College, Augustana College, Illinois
- Don Boileau, Professor of Communication, George Mason University, Virginia
- Elizabeth Ciner, Associate Dean of the College, Carleton College, Minnesota
- Ariane Hoy, Senior Program Officer, Bonner Foundation, New Jersey

- Katherine Lang, Chair, History Department, University of Wisconsin-Eau Claire
- Adam Lutzker, Associate Professor and Chair, Economics, The University of Michigan-Flint
- Jean Mach, Professor of English, College of San Mateo, California
- Marcia Mentkowski, Senior Scholar for Educational Research, Alverno College, Wisconsin
- Francine Navakas, Bramsen Professor in the Humanities; Associate Academic Dean, North Central College, Illinois
- Judy Patton, Associate Dean of the School of Fine and Performing Arts and Professor of Theater Art, Portland State University, Oregon
- Candyce Reynolds, Faculty Member, Educational Leadership and Policy, Portland State University
- William Rickards, Senior Research Associate, Educational Research and Evaluation, Alverno College, Wisconsin
- Judith Stanley, Professor of English, Alverno College, Wisconsin

Creative Thinking Rubric
- Dorothy Keyser, Associate Professor, Music, The University of North Dakota
- Patrice Caldwell, Executive Director of Planning and Analysis, Eastern New Mexico University
- Theresa Ford, Director of Educational Assessment, The College of Wooster, Ohio
- Stephanie Gibson, Associate Professor, School of Communications Design, The University of Baltimore, Maryland
- Patrick McGovern, Director of Membership Development, Acacia Fraternity International Headquarters, Indiana
- Shirley Keeton, Coordinator of Institutional Research; Assistant Professor of Sociology, Purdue University North Central, Indiana
- Nancy Grace Professor of English The College of Wooster, Ohio

Written Communications Rubric

- Linda Adler-Kassner, Professor and Director of First-Year Writing, Eastern Michigan University
- Theresa Flateby, Director of University Assessment, Evaluation and Testing, University of South Florida
- Susanmarie Harrington, Director of Writing in the Disciplines, University of Vermont
- Jean Mach, Professor of English, College of San Mateo
- Noreen O'Connor, Assistant Professor of English, King's College, Pennsylvania
- Carol Rutz, Director, Writing Program, Carleton College, Minnesota

Teamwork Rubric

- Tina Clawson, Associate Director, Outreach, Oregon State University
- Taz Daughtrey, Instructor, Computer Science, James Madison University, Virginia
- Rolf Enger, Director of Education, United States Air Force Academy, Colorado
- Steven Jones, Director of Academic Assessment, United States Air Force Academy, Colorado
- Richard Hughes, USAF Academy Transformation Chair, United States Air Force Academy
- Lynne Mason, Associate Professor, School of Applied and Information Technology, Community College of Baltimore County-Catonsville, Maryland
- Nancy O'Laughlin, IT-Client Support and Services, University of Delaware
- Kathleen Pusecker, Associate Director, Office of Educational Assessment, University of Delaware
- Kimberly Thompson, Director of Assessment, Regis University, Colorado

Quantitative Literacy Rubric

- Michael Burke, Professor of Mathematics, College of San Mateo, California
- Rolf Enger, Director of Education, United States Air Force Academy, Colorado
- Nathan Grawe, Associate Professor of Economics, Carleton College, Minnesota
- Joan Hawthorne, Assistant Provost for Assessment, University of North Dakota
- Richard Hughes, USAF Academy Transformation Chair, United States Air Force Academy
- Steven Jones, Director of Academic Assessment, United States Air Force Academy, Colorado
- Jean Mach, Professor of English, College of San Mateo, California
- Corrine Taylor, Director, Lee Day Gillespie '49 Quantitative Reasoning Program, Wellesley College, Massachusetts

Critical Thinking Rubric

- Gregory Basshan, Chair and Professor of Philosophy, King's College, Pennsylvania
- Gary Brown, Director, The Center for Teaching, Learning, and Technology, Washington State University
- Sandy Figueroa, Assistant Professor, Computer Information Systems and Technology, Hostos Community College, New York
- R. Johnson, Director of Instructional Development and Research, The Pennsylvania State University
- Jean Mach, Professor of English, College of San Mateo, California
- Jean O'Brien, Professor of Psychology, King's College, Pennsylvania
- Tanya Renner, Professor of Psychology, Kapi'olani Community College, Hawaii
- Mary Walczak, Associate Professor and Chair of Chemistry; Director of Evaluation and Assessment, St. Olaf College, Wisconsin

Inquiry and Analysis Rubric

- Lea Campbell, Director of Academic Assessment, University of Houston-Downtown, Texas
- Kathryne McConnell, University Academic Assessment Coordinator, Virginia Polytechnic Institute and State University
- Michael Greene, Coordinator of Baccalaureate Programs, Cayuga Community College, New York
- Milton Hakel, Professor and Ohio Eminent Scholar, Bowling Green State University, Ohio
- Anne Herrington, Professor of English, University of Massachusetts Amherst
- Robin Jeffers, Coordinator, Outcomes Assessment/Institutional Effectiveness, Bellevue Community College, Washington
- Jessica Jonson, Director of Institutional Assessment, University of Nebraska-Lincoln
- Jacqulyn Laeur-Glebov, Associate Director of Institutional Research & Assessment, Carleton College, Minnesota
- Cornelia Paraskevas, Professor, Linguistics, Western Oregon University

Information Literacy Rubric

- James Dutt, Director, Center for Excellence in Teaching and Learning, The University of Baltimore, Maryland
- Elizabeth Knapik, Director of Information Literacy Programs, Sacred Heart University, Connecticut
- Andrew Marx, Assistant Professor, Core Curriculum, Virginia Commonwealth University
- Terrence Mech, Director of the Library, King's College, Pennsylvania
- Megan Oakleaf, Assistant Professor and Dean, Information Studies, Syracuse University, New York
- Gretchen Sauvey, United States Institute of Peace, Washington, D.C.
- Debbie Schwartz, Associate Dean of Institutional Assessment, Lourdes College, Ohio

- Wilbur Stolt, Director of Libraries, University of North Dakota
- Anne Zald, Head of Instruction, University of Nevada Las Vegas

Foundations and Skills for Lifelong Learning Rubric:
- Debra Buchanan, Assistant Vice President for Academic Affairs; Associate Provost, Jackson State University, Mississippi
- Keston Fulcher, Assistant Professor of Graduate Psychology and Associate Assessment Specialist, James Madison University, Virginia
- Lynne Groves, Director, Instructional Strategies, Minnesota State Colleges and Universities
- Rose Mince, Dean of Instruction for Curriculum and Assessment, Community College of Baltimore County-Essex
- Mary Somerville, University Librarian/Director, University of Colorado Denver
- Suzanne Weinstein, Manager of Instructional Consulting and Coordinator of Academic Assessment, Pennsylvania State University
- Judith Wertheim, Vice President, Higher Education Services, Council for Adult and Experiential Learning, Illinois

Problem Solving Rubric
- John Bennett, Emeritus Associate Dean and Associate Professor, University of Connecticut
- Avon Chapman, Director, Adjunct Development and Faculty Administrative Support Services, Atlantic Cape Community College, New Jersey
- Kathy Faggiani, Milwaukee School of Engineering, Wisconsin
- Heidi Fencl, Associate Professor and Chair of Physics, University of Wisconsin-Green Bay
- Nancy Mattina, Faculty and Director, Adult Degree Program, Prescott College, Arizona
- William Murry, Director of Institutional Assessment, University of San Francisco, California

- Joni Spurlin, University Director of Assessment and Associate Director of University Planning and Analysis, North Carolina State University
- Pamela Steinke, Director of Research, Planning and Assessment, Meredith College, North Carolina
- Jannis Taylor, Learning Services Coordinator, Maryville College, Tennessee

Ethical Reasoning Rubric

- Alan Belcher, Assistant to the Provost, University of Charleston, South Carolina
- Beth Dyer, Manager, Administration and Technological Services, Oregon State University
- Amy Gort, Dean, College of Arts and Sciences, Concordia University, Oregon
- Lou Matz, Associate Dean and Director of General Education and Associate Professor of Philosophy, University of the Pacific
- Nancy Mitchell, Interim Director for General Education; Professor of Advertising, University of Nebraska-Lincoln
- Eric Moore, Assistant Professor of Philosophy, Longwood University, Virginia
- Jane Wilson, Dean of Fine Arts, Assessment and Professional Development, North Hennepin Community College, Minnesota

Civic Engagement Rubric

- Leila Brammer, Associate Professor and Chair of Communication Studies, Gustavus Adolphus College, Minnesota
- Julie Hatcher, Associate Director, Center for Service and Learning, Indiana University-Purdue University Indianapolis
- Donna Gessell, Executive Director for Regional Engagement, Professor of English North, Georgia College and State University
- Ariane Hoy, Senior Program Officer, Bonner Foundation, New Jersey
- Mary Kirlin, Associate Professor, Public Policy & Administration, California State University, Sacramento

- Kathleen Weigert, Executive Director, Center for Social Justice Research, Teaching and Service; Research Professor of Sociology and Anthropology and Program on Justice and Peace, Georgetown University, Washington, D.C.
- Lori Muntz, English Program Coordinator, Iowa Wesleyan College
- Susana Rivera-Mills, Chair of Foreign Languages and Literatures, Associate Professor of Spanish and Diversity Advancement, Oregon State University
- John Saltmarsh, Director of New England Resource Center for Higher Education, University of Massachusetts Boston
- Amy Spring, Assistant Director, Community-University Partnerships Center for Academic Excellence, Portland State University, Oregon
- Jean Strait, Associate Professor of Education, Hamline University, Minnesota

Intercultural Knowledge and Competence Rubric
- Janet Bennett, Executive Director, The Intercultural Communication Institute, Oregon
- Kimberly Brown, Professor of Applied Linguistics, Portland State University, Oregon
- Chris Cartwright, Doctoral Student, Educational Leadership, Portland State University, Oregon
- Margaret Davis, Professor Altmayer Chair of Literature; Director of Core Curriculum, Spring Hill College, Alabama
- Darla Deardorff, Executive Director, AIEA, Duke University, North Carolina
- Debbin Gin, Director, Diversity Studies, Azusa Pacific University, California
- Carole Huston, Assessment Director and Professor of Communication Studies, University of San Diego, California
- Lee Knefelkamp, Professor of Psychology and Education, Teachers College, Columbia University, New York
- Masami Nishishiba, Assistant Professor, Public Administration-Urban and Public Affairs, Portland State University, Oregon

- Daryl Smith, Professor of Education and Psychology, Claremont Graduate University

Field-Test e-Portfolio Campuses
- Alverno College, Wisconsin
- Bowling Green State University, Ohio
- City University of New York-LaGuardia
- Community College College of San Mateo, California
- George Mason University, Virginia
- Kapi'olani Community College, Hawaii
- Portland State University, Oregon
- Rose-Hulman Institute of Technology, Indiana
- San Francisco State University, California
- Spelman College, Georgia
- St. Olaf College, Minnesota
- University of Michigan

AASCU-led Degrees of Preparation Survey Project

Experts Panel
- Jutta Birmelle, Professor, California State University-Long Beach
- Andrew Downs, Assistant Professor, Indiana University Purdue University, Fort Wayne
- Gerald Eisman, Professor, San Francisco State University, California
- Constance Flanagan, Professor, Pennsylvania State University
- Susan Peters, Associate Professor and Department Chair, California Polytechnic University - Pomona
- Amy Kay Syvertsen, Graduate Intern, Pennsylvania State University
- Roberta Teahen, Associate Vice President for Academic Affairs, Ferris State University, Michigan
- Michael Wolf, Assistant Professor, Indiana University Purdue University, Fort Wayne

Validation Team
- James David Ballard, Professor of Sociology, California State University, Northridge
- Katie Busby, Director of Student Affairs Assessment and Planning, University of Alabama
- Elizabeth Creamer, Professor, Education Leadership and Policy, Virginia Tech
- J.E. (Ernie) Gonzalez, Director of Institutional Research, University of Southern Florida, St. Petersburg
- Marsha Hirano-Nakanishi, Assistant Vice Chancellor, Academic Research, California State University
- Bettina Huber, Director of Institutional Research, California State University, Northridge
- Judith Ouimet, Professor, Department of Education, University of Nevada, Reno
- Gary Pike, Executive Director, Information Management and Institutional Research, Indiana University Purdue University Indianapolis
- John Pryor, Director, Cooperative Institutional Research Program, University of California-Los Angeles (CIRP)

Field-Test Institutions
- Berea College, Kentucky
- California State University, Fullerton
- California State University, Northridge
- Ferris State University, Michigan
- Fitchburg State College, Massachusetts
- Hampshire College, Massachusetts
- Indiana University Purdue University, Indianapolis
- Northern Arizona University
- Prairie View A&M University, Texas
- San Francisco State University, California
- Smith College, Massachusetts
- South Dakota State University
- University of Nevada, Reno
- University of Wisconsin-Parkside

APLU-led Test Validity Study of VSA Learning Outcomes Instruments

Researchers Primary Authors
- Stephen Klein, Council for Aid to Education, New York
- Ou Lydia Liu, Educational Testing Service, New Jersey
- James Sconing, ACT, Iowa

Secondary Authors
- Roger Bolus, Council for Aid to Education, New York
- Brent Bridgeman, Educational Testing Service, New Jersey
- Heather Kuglemass, Council for Aid to Education, New York
- Alexander Nemeth, Council for Aid to Education, New York
- Steven Robbins, ACT, Iowa
- Jeffrey Streedle, Council for Aid to Education, New York

Field-Test Institutions
- Alabama A&M University
- Arizona State University, Tempe Campus,
- Boise State University, Idaho
- California State University, Northridge
- Florida State University
- Kalamazoo College, Michian
- Massachusetts Institute of Technology
- University of Colorado at Denver
- University of Michigan-Ann Arbor
- University of Minnesota-Twin Cities
- University of Texas at El Paso
- University of Vermont
- University of Wisconsin-Stout