Measuring Core Educational Outcomes at Research Universities for Improvement and Accountability

Prepared for the October 2008 seminar on Measuring Undergraduate Learning Outcomes: A Working Agenda for Public Research Universities
University of Minnesotta

David Shulenburger
Vice President for Academic Affairs
National Association of State Colleges and Land-Grant Universities

Measuring Core Educational Outcomes at Research Universities for Improvement and Accountability

Abstract: The new public expectation that universities measure core educational outcomes and report the results of those measurements publicly has perturbed what had become a fairly stable assessment ecology on public university campuses. This paper examines the roots of the new demands, contrasts value-added core learning outcome measurement with the measurements that flow from existing assessment programs and works through the issues with which the academy must deal if measurement of core educational outcomes is to satisfy the public demand for accountability while providing information with real utility for improving higher education. Much can be gained if we undertake core outcomes measurement willingly and work to improve its usefulness to the academy while ensuring that it complements, not replaces, existing assessment programs.

After expressing considerable dissatisfaction with transparency of higher education, the National Commission on the Future of Higher Education in fall, 2006 recommended:

> *"Colleges and universities must become more transparent about cost, price, and student success outcomes, and must willingly share this information with students and families. Student achievement, which is inextricably connected to institutional success, must be measured by institutions on a "value-added" basis that takes into account students' academic baselines when assessing their results. This information should be made available to students, and reported publicly in aggregate form to provide consumers and policymakers an accessible, understandable way to measure the relative effectiveness of different colleges and universities".*[1]

Writing earlier in the same year in his insightful book Our Underachieving Colleges, Derek Bok observed,

> "*Professors seldom receive clear evidence of how much their students are learning*," [2]

And he broadened his purview from the classroom to the undergraduate education enterprise when he observed,

[1] A Test of Leadership, Charting the Future of U.S. Higher Education, The Commission on the Future of Higher Education, A Report of the Commission Appointed by Secretary of Education Margaret Spellings, September 2006, p.4.

[2] Our Underachieving Colleges, Derek Bok, Princeton University Press, 2006. p. 315

> *". . . in sharp contrast to many human endeavors no one knows whether college students are writing better or thinking more rigorously or making more progress toward other educational goals than they were 50 years ago."*[3]

Finally, Bok advocated a fix,

> *". . . yet many other pertinent competencies do lend themselves to rough but useable assessments. Student writing, quantitative reasoning, foreign language proficiency, and critical thinking and analysis are all pertinent examples. Using available measures, colleges can discover important skills that are not being measured, subjects students do not truly understand, entire groups of undergraduates that are performing well below their apparent capabilities. With these problems identified, faculties can experiment with these new methods of teaching, and gradually develop more effective ways of improving student learning that can be described and used successfully by others. It is the failure of most college faculties to make serious efforts of this kind that merits disappointment over the way undergraduate education has evolved."*[4]

Strangely, both the National Commission and President Bok's conclusions were drawn twenty-five years after the dawning of what Peter Ewell has dubbed the "age of accountability."[5] This "age" is defined by essentially every university and every accredited program within them having developed and implemented assessment programs, sometimes of their own volition, but more often in response to a requirement from their institutional or specialized accreditor that they do so.

To this "age of accountability" the Spelling's Commission gave short shrift:

> *"Despite increased attention to student learning results by colleges and universities and accreditation agencies, parents and students have no solid evidence, comparable across institutions of how much students learn in colleges or whether they learn more at one college than another. Similarly, policymakers need more comprehensive data to help them decide whether the national investment in higher education is paying off and how tax payer dollars could be used more effectively."* [6]

The Commission made much of the findings of the National Adult Literacy Survey which they characterized as indicating that between 1992 and 2003 the proportion of college graduates proficient in prose literacy decreased for all levels by over 20% and those proficient in document literacy declined by over 30%.[7] While the Commission was mistaken in interpreting this data as evidence that higher education was doing a

[3] Ibid. 318
[4] Ibid. 321
[5] U.S. Accreditation and the Future of Quality Assurance, prepared by Peter T. Ewell, CHEA, Washington D.C. 2008 pp. 42 and 43.
[6] Op. cit. p. 14.
[7] Ibid. p.14.

worse job of educating students,[8] its skepticism about the performance of colleges was illustrative of a national pattern.

From 1987 to 1997 Margaret Miller, was the Associate Director of Academic Affairs for the State Council of Higher Education for Virginia. She now serves as a professor in the Curry School of Education of the University of Virginia and is editor of *Change Magazine*. During her decade in the Academic Affairs position she annually reviewed the assessment reports from the many public universities in the state of Virginia. Her observation about those reports was that she found nearly every one of them impressive and filled with important results, but that there was no way she could conclude from them whether learning had improved over the decade at any of those universities. Every year was a new year so far as the findings were concerned.

So what is it about methods of assessment currently used that permitted the Commission (and Margaret Miller) to dismiss them as not providing ". . . solid evidence, comparable across institutions of how much students learn in colleges or whether they learn more at one college than another?" I suggest there are three things:

1) The results of the assessments we now conduct are not generally made public so declarations by university officials that student learning is at a satisfactory level or that student learning is improving is supported only by "trust me" statements, not data.
2) The assessment instruments in use are varied and often are changed from one assessment to the next. Many instruments are developed by faculty in a school or department and are unique to a single unit while others are standardized.
3) What is measured by assessment programs generally relates to portions of the educational program, e.g. general education or the major, and not to the results of the entire educational process.

What differentiates this call for assessment from earlier ones is specificity. It is a call for transparent and comparable evaluation of the education the whole student receives, not of the individual elements that go into making up the student's education. On the latter point, an apt analogy to the "age of assessment" would be to an auto assembly line on which every part is tested thoroughly before assembly, but the assembled car is not started and driven off the line. What Bok and the Spellings Commission call for is for us to start that car and drive it away from the line. The new age of assessment also specifically enumerates the types of testing of the finished product, the graduate, they wish for higher education to conduct. Critical thinking, problem solving and written communications appear on the list. To continue the auto

[8] There are two reasons for this statement. First, the sample of college graduates was a national sample of all those who had graduated from college, not of recent college graduates. Subsequent administrations of the NALS to populations made up only of recent college graduates did not discover any decline in literacy. In addition, the National Research Council in its review of the NALS warned the Department of Education that "the NALS will not allow for detection of problems at the level of proficiency" (Measuring Literacy, National Research Council, p. 166), unambiguously arguing that the test as constructed simply could not detect who was "proficient" and who was not.

assembly analogy, what they call for is analogous to testing the assembled automobile on performance dimensions like maneuverability, acceleration and gas mileage.

This is not to suggest that continued assessment of general education and education in the major is unnecessary; it is needed for the same reason that testing of the automobile's components is needed: the whole is unlikely to work if the component parts are faulty. Unfortunately, working parts do not necessarily produce a working whole. Both general and major assessments are needed if we are to meet general education goals and if the specialized education in the major is to prepare students to practice professions. This new demand is for testing of the whole student on selected core educational outcome dimensions, not for abandoning assessment of general education and the major.

Testing of the Whole Student is Desired: Some in the academy have argued that testing of graduates on the core educational outcomes of critical thinking, problem solving and written communication has to be done along disciplinary lines.[9] [10] The argument is that a physicist, for example, goes about structuring thought, problem-solving and writing up conclusions in a different manner from those studying other disciplines. This is undoubtedly true, but the analysis fails to take into account the many expressions of concern by policy makers, employers, board members and others that all college graduates should possess skills of critical thinking, problem-solving and written communication that will be used in a wide array of situations, not just those that relate to their majors in college. Today's knowledge-based economy places less emphasis on acquiring content knowledge and more on evaluating data and sources, solving problems, and effectively communicating information to a variety of audiences.

The "stove-piping" that occurs within the major is the university's way of ensuring that students thoroughly learn a disciplinary subject. However, when students leave the university, relatively few enter work that directly relates to their majors. Even those who do enter discipline-connected employment generally find that over time they are asked to take on broader responsibilities by their employers. Thus while testing of students within the major is important to ensure that education in the major is effective, testing of critical thinking, problem solving and written communications *across all majors* is an essential element in determining whether our graduates are really ready for the broad set of situations to which they will apply their skills after graduation.

We increase the tendency to focus excessively on majors when we conceive of skills like critical thinking, problem-solving and written communications as arising only out of general education. If the student's work in the major does not build on the foundation of core skills acquired in the first two years of university study, those skills are likely to atrophy, not grow. In the development of writing skills, for example, we have learned that when writing is stressed throughout the curriculum, not solely in the first year composition courses, much more progress occurs. Assigning the development

[9] What's wrong with the VSA Approach? Joan Hawthorne, *Liberal Education* (Spring 2008), p. 27.

[10] **Institutional Versus Academic Discipline Measures of Student Experience: A Matter of Relative Validity.** Steve Chatman. CSHE.8.07. (May 2007)

of critical thinking and problem-solving skills to the exclusive domain of "general education" is to suggest that faculty members teaching in the major have no responsibility for reinforcing and enhancing those skills. Giving such a pass to disciplinary faculty serves our students poorly.

Practical Problems in Core Educational Outcome Assessment

Concluding that we should test graduates on the core educational outcomes of critical thinking, problem solving and written communications leaves several practical problems. Among them are:

1- What tests are available to measure these high level core educational outcomes?
2- Are the tests "good" measures of these core educational outcomes?
3- What cut-off scores signify satisfactory minimum attainment?
4- How do we use the results of value-added core educational outcomes testing for curricular improvement?
5- How do universities afford yet more testing?

1- What tests are available to measure these high level core educational outcomes? In the process of developing the Voluntary System of Accountability, NASULGC and AASCU received strongly felt advice (particularly from its provosts groups) that multiple instrument options should be made available for measuring core educational outcomes. That advice reflected the concern that use of a single measurement instrument might have the subtle effect of reducing the diversity of approaches member universities used in their educational programs. Those expressing concern often referenced the effect that No Child Left Behind testing appears to be having on diversity of curricular offerings in public schools and argued that use of multiple tests would mitigate this effect.[11]

To select the tests to be used in the Voluntary System of Accountability NASULGC and AASCU assembled both a working group and a taskforce constituted of university presidents, provosts, and specialists. Their first task was to identify criteria to evaluate tests. They decided upon the following criteria set:

- Usability to measure value-added
- Includes measures of critical thinking, analytical reasoning, and written communication

[11] Since there is no single national test (individual states select their own test), this fear was somewhat off target. On the other hand, the NCLB tests measure the same basic skills, e.g. reading and mathematics, and their use has apparently reduced curricular emphasis on subjects like social studies and art that are not included in the testing. Unlike NCLB's relatively low level test of skills, the high level core outcomes tests are independent of subject matter and, therefore, should not have the effect of narrowing the curricula of universities using them.

- Can be used for "benchmarking" or appropriate comparisons from normed groups or criteria.
- Both valid and reliable
- Results can be used to improve teaching and learning
- Appropriate for institutions with diverse missions, student bodies, and admission criteria.

The groups assembled the following fourteen candidate tests and measured them against the criteria set:

Collegiate Assessment of Academic Proficiency (CAAP)
College Basic Academic Subjects Examination (C-Base)
Collegiate Learning Assessment (CLA)
College Level Academic Skills Test
Electronic Portfolios[12]
Group Assessment of Logical Thinking
Graduate Record Examination (GRE)
Georgia Regents Test
Information and Communication Technology Literacy Assessment
Measurement of Academic Proficiency & Progress (MAPP)
National Assessment of Adult Literacy
National Survey of Student Engagement
Standardized Assessment of Information Literacy Skills
ACT WorkKeys

Based on the review the group selected six tests (CAAP, MAPP, C-Base, GRE, CLA and WorkKeys) for more thorough evaluation and solicited detailed information from the test makers. (The questionnaire used to solicit that information appears in the appendix to this paper.) Based upon the questionnaire responses the groups narrowed the set of tests that best met the criteria to GRE, CAAP, MAPP and CLA. [13] Based on the evaluation and recommendations of the workgroup and the taskforce, the VSA presidential advisory board ultimately found that CAAP, MAPP and CLA each sufficiently met the criteria listed above and decided that each of them (subject to confirmation that CAAP and MAPP each could be used in the same fashion as CLA for providing comparative value-added data) should be made available as alternative core outcomes evaluation exams in VSA.

[12] Electronic Portfolios, unlike the other instruments listed is not a discrete test of core educational outcomes. Rather it is a category that includes many collections of student work at various universities. While the workgroup considered the data in portfolios to be an attractive source for evaluating core educational outcomes, the lack of a nationally agreed-upon set of rubrics to evaluate core educational outcomes forced the group to exclude portfolios from subsequent evaluation as it failed to meet the comparability criterion. AAC&U is currently developing such rubrics under a FIPSE grant.

[13] GRE and Educational Testing Service decided subsequent to the evaluation by the work group and taskforce that they could not support the use of GRE as a core educational evaluation instrument and withdrew it from further consideration for use in VSA.

2- Are the tests "good" measures of core educational outcomes? This is a difficult question with many facets.

First, a bit about the three tests: **CAAP and MAPP**, as used to measure the dimensions of critical thinking, problem solving and written communications,[14] are made of batteries of objective questions (true/false, multiple choice) used to measure the critical thinking/problem solving dimensions and a written essay section to judge written communications ability. Both tests produce multiple subscores and, if used with optional general education components, can produce an array of subscores on general education subject matter mastery in addition to core educational outcomes. A principal difference between them is that CAAP is administered in a paper/pencil format and MAPP is administered in an electronic format.

CLA is an all-essay response exam administered in an electronic format. Students are given written prompts and provided access to reports, data, new articles, memos, etc., that they may consult as they write their answers. Students complete a performance task and an analytic writing task with two components, break an argument and make an argument. Subscores are reported for each of the exercises as well as an overall score. Unlike CAAP and MAPP, CLA does not report separate critical thinking and written communications scores as they argue that written communications ability is so closely intertwined with thinking and reasoning abilities that production of separate scores is a meaningless task.[15]

The three tests have certain similarities:

They are each core education measurements, not general education measurements. As they are used in VSA, it is simply incorrect to call any of the three tests sets "general education" tests. While they measure skills that might be developed in general education, those skills ought to be honed throughout the undergraduate's experience.

Each of testing agencies has agreed to use the following identical technique and conventions in reporting value-added scores:

> *Each testing company will employ Ordinary Least Squares (OLS) regression models that use--at the school level--mean CLA, CAAP or MAPP scores as the dependent variable and mean SAT or ACT scores as the independent variable. If a participating institution does not collect SAT (ACT) scores, ACT (SAT) scores are converted to the SAT's (ACT's) scale of measurement using a standard cross-walk. One equation will be run using freshmen's test scores and another equation using seniors' test scores. Both equations will use the institution as the unit of analysis. These equations establish the typical relationship between a*

[14] Actually, ACT, the maker of CAAP, holds that there are really just two dimensions, critical thinking and written communication, as problem solving is not a separate dimension but a component of critical thinking.

[15] Holistic Tests in a Sub-score World: The Diagnostic Logic of the CLA, Roger Benjamin, Marc Chun and Richard Shavelson, CAE, 2007 http://www.cae.org/content/pdf/WhitePaperHolisticTests.pdf

school's mean test score and the average academic ability (as measured by SAT (ACT) performance) of its incoming students who participate in the testing program. These relationships will be used to predict a school's mean test score. Value-added scores are based on the differences between a school's actual test score and its predicted test score.

Differences between actual and expected scores are reported in two ways: (1) "points" on the test scale and (2) standard errors. The latter is used to facilitate comparisons and define the performance levels as follows. Colleges with actual mean scores between -1.00 to +1.00 standard errors from their expected scores are categorized as being At Expected. Institutions with actual mean test scores greater than one standard error (but less than two standard errors) from their expected scores are in the Above Expected or Below Expected categories (depending on the direction of the deviation). The schools with actual scores greater than two standard errors from their expected scores are in the Well Above Expected or Well Below Expected categories.

There are actual and potential differences among the tests.

Test Validity simply reflects whether the exam measures what it was constructed to measure. In the case of critical thinking, demonstration of validity in a mechanical fashion is difficult as the definition of critical thinking is necessarily somewhat fuzzy. ETS (MAPP), ACT (CAPP) and CAE (CLA) each claim that their tests are valid in that they measure critical thinking but the broad and amorphous nature of the definition of critical thinking necessarily leaves opportunity for debate. This contrasts with, for example, the validity of a test to determine if subjects can do simple arithmetic. Simple arithmetic's computations are rigid and precise and the validity of the test can be cleanly established.

One of the differences among the three tests relates to a facet of validity often called "face" validity. Face validity is simply the judgment of those who know the field as to whether a test measures those phenomena. The CLA gets very high marks for face validity. Its prompts are similar to situations one might encounter in the professional work place. The data made available as a resource to the test taker is of the sort one commonly encounters when working through a work place problem. Scoring is based on the degree to which the student uses the available data in a reasoned fashion in reaching a decision and the ability of the student to articulate both the rationale and the decision. Those who have worked in business, government or education settings readily express the opinion that the CLA has face validity. Face validity is bit more difficult to declare for CAAP and MAPP because the limited response format of true/false and multiple choice has few counterparts on the job, but their prompts are consistent with strong face validity.

A second kind of validity is termed "concurrent" validity [16]. Concurrent validity is high if one test of phenomena has scores that correlate positively at a very

[16] For a simple presentation on validity and reliability that informed this discussion see Research Methods: Variables, Validity and Reliability at http://allpsych.com/researchmethods/validityreliability.html

high level with scores on another test of the same phenomena. In this case, the concurrent validity question is whether CLA, CAAP and MAPP measure the same thing. During fall semester 2008 over 1200 students at 13 different universities will take each of these three tests in a construct validity study funded by FIPSE. That study, the key questions it will answer and the specifics of the sample and methodology are described in Appendix 2 of this paper.

Another type of validity is "content" validity, the ability of a test to represent all of the content within a particular construct. Since the construct here is defined by the content of tests, i.e., critical thinking, problem solving and written communication, the three tests score high on content validity. However if one intention is to measure all of the content of an undergraduate education, the three tests have low content validity. Here, it is important to claim to measure only what the tests are intended to measure. For many reasons, the use of CAAP, MAPP or CLA or value-added scores derived from them to rank undergraduate programs or to make funding decisions concerning them is inappropriate. Given the limited set of content they measure and the enormous variety of content contained in the numerous undergraduate programs, an overall undergraduate program ranking based on scores from these tests would be particularly meaningless.

A final note on validity is important. Both the CLA and MAPP are in electronic format but CAAP is in pen and paper format. Does the form in which a test is administered affect what it measures? The FIPSE-funded construct validity study will shed some light on this question.

Reliability is the ability to repeat the test and get roughly the same score. There has been controversy about reliability, particularly regarding the CLA. Trudy Banta has argued that the value-added measurement is inherently unreliable with very low test-retest correlation (r = .1) across individuals.[17] Steven Klein, et. al., countered that test-retest reliability for value-added testing across groups of individuals is high (r = .7 to .77).[18] Banta's claim is essentially that the variance in measuring outcomes for freshmen when compounded by the variance in measuring outcomes for seniors makes any attempt to use the difference between them as a measure of gain statistically meaningless. Klein et. al. countered that when the subject of inquiry is differences in test scores for *groups of students*, not individual students, gain can be measured with considerable reliability. Ultimately the eminent psychometrican, Howard Wainer sorted out the differences. He sided with Klein et. al. citing recent research findings on the subject to support his conclusions.[19] Thus it is not a question of whether Banta or Klein was correct for both were correct. The question is the use to which the test is put.

---

[17] *A Warning on Measuring Learning Outcomes***,** Trudy Banta, Inside Higher Education, January 26, 2007.
[18] *Setting the Record Straight,* Stephen Klein, Richard Shavelson and Roger Benjamin. February 8, 2007
[19] For summaries of the research see Chapter 26 in the Handbook of Statistics (Volume 27) Psychometrics (Eds C. R. Rao and S.Sinharay). Elsevier Science, Amsterdam. 2007, pages 893-918 and
Value-Added Assessment. In Handbook of Statistics (Volume 27) Psychometrics (Eds C. R. Rao and S. Sinharay). Elsevier Science, Amsterdam. 2007, pages 867-892

Finally, for research universities a good measure of value-added core educational outcomes must meet one additional criterion: it must be equally as effective in measuring learning gains for them as for non- research universities with their generally lower admissions standards. This matter has caused considerable concern at some highly selective institutions. They reason that because their students have already achieved in high school much more than those admitted to less selective universities, the latter group of students has much more to learn in college than does their group of new students. Therefore learning gains for their students are likely to be lower than those for students at the less selective university. Alternately, they reason that instruments that measure learning gains might not be sophisticated or sensitive enough to measure all the gain they impart to their students, i.e. the test may not be challenging enough. Either way, they fear that they will be relatively disadvantaged by learning gain measurement.

The empirical data that we have is not consistent with this logic. Appendix III displays graphs from the administrations of CLA in 2005, 2006 and 2007. In each graph both freshmen scores and senior scores are directly related to the ACT or SATs of the students tested. Regressions show the relationships to be linear. The intercepts for the senior's regression lines are significantly above those of the freshmen's regression lines with the slope coefficients being of roughly the same magnitude. For each of these three years the CLA has found that the magnitude of the learning gains is unrelated to the ACT/SAT scores of incoming students.

CLA is designed taking care that neither freshmen nor seniors "top out" with any great frequency on the instrument. Moreover, while the percent of freshmen or seniors with maximum scores on the CLA was very small prior to fall 2007, in fall 2007, CLA ceased the practice of capping scores. Raw scores on each of the tasks are now converted to a common scale of measurement with the same mean and standard deviation as the SAT score of the freshmen who took the SAT. While this procedure normally results in the CLA scores falling within the SAT's range of 400 to 1600, on very rare occasions when a student achieves a score well above or below that of the other students taking a task, a scale score is assigned that is outside the normal SAT range. Thus, there is no minimum or maximum score on the CLA and no "topping out" problem.

Thus, there is room to demonstrate learning gain and, as Appendix III illustrates, students at selective universities demonstrate that gain. Over more than 30 years I had the privilege to work with a large number of students who came to the University of Kansas with ACT/SAT scores very near the maximum level possible. I watched them grow intellectually over their years at the university and saw them emerge from the university as graduates much more polished and eloquent than when I first met them. The learning gain I've seen in these very bright students is what the CLA attempts to measure. I do not believe that any highly selective university actually believes that their students do not learn and grow intellectually while they are enrolled. The question is whether the testing instrument, i.e. CLA, CAAP or MAPP, is capable of measuring that gain.

Remember also that the statistical technique for measuring learning gain used by the CLA, the expected value calculation, controls for the ability of entering students. Thus a given university's value-added is contrasted with what would have been expected given the ability of its entering students, not with the value- added of a university that admitted less able students. The same technique will be used by CAAP and MAPP to calculate value added at universities participating in VSA.

Will CAAP and MAPP, with their multiple-choice, true-false and fill in the blank formats be able to measure learning gain equally well at open admissions and highly selective universities? I believe this to be an empirical rather than a theoretical question. Surely one with enough skill as a test maker can develop questions of such nuance that intellect of the highest ability will be required to arrive at the correct answer. The continued existence of crossword puzzle contests and chess matches surely illustrates that even in situations when response options are limited, differential levels of skill and ability can be distinguished.

The FIPSE construct validity study includes some very highly selective institutions and will provide data like that in Appendix III for CAAP and MAPP and will supply additional data for CLA. Other highly selective institutions are participants in VSA and may choose to administer CAAP or MAPP during the four-year trial period. Thus empirical evidence will be forthcoming for these two tests and data will continue to be amassed for the CLA at universities of all selectivity levels.

Finally, the academy itself has to determine how good these tests are. The data produced by the FIPSE construct validity study will inform our colleagues. But perhaps a more important test will be the program of trial administrations of the three tests that the more than 250 universities who have signed onto VSA have agreed to conduct over the next four years. Whether the tests are "good" will be determined by practice and institutional use of results.

3- What cut-off scores signify satisfactory minimum attainment? This is indeed an important question. The tests themselves measure only the level of attainment. How much achievement must be documented for the award of a bachelor's degree? This is a question that neither the U.S. Department of Education, nor the academy, nor governing boards, nor state regulators have answered.

We have no agreement on the minimum level of educational content or attainment that should constitute a bachelor's degree. While that seems odd, it follows from the fact that we also have no agreement on what minimum level of educational attainment should be required for award of a high school degree. In addition, some universities choose to be highly selective and do not admit all high school graduates. U.S. universities therefore admit students with widely differing levels of educational attainment as freshmen, making it difficult for universities to move all admitted freshmen to the same minimum end point after a four-year undergraduate experience.

It is also true that the U.S. has no Ministry of Education empowered to set such standards. We value the diversity produced by our higher education system and we value the independence enjoyed by each university that permits diversity to flourish. We resist efforts by the U.S. Department of Education or Congress to establish such standards. I note that the Bologna Process within the European Community is slowly chipping away at diversity among its universities. They are working toward being able to specify the core of what a graduate with a given degree should know and be able to do. The Bologna Process has been pushed along as the mismatch between physically small nations and physically broader labor markets increasingly limited employment opportunities for those with degree types unique to their nation. In the United States we have been more tolerant of extreme diversity but Friedman's "flat world" rapidly may force us into a Bologna-type activity.

Core educational outcomes test scores correlate positively at very high levels (.8+) with admission test scores such as ACT and SAT. Were some threshold raw score on core educational outcomes tests to be used as a cut score in core educational outcomes measurement, the degree to which universities succeed in achieving that threshold on CAAP, MAPP or CLA probably would be primarily a direct reflection of the admissions selectivity of those universities. Given the very wide range of entering freshmen test scores and grades, meeting a threshold on an outcomes test at graduation is unlikely to reflect the degree to which the differing university environments foster differences in learning among students.

The surest way and the most resource effective way for a university to increase the proportion of its students meeting any pre established outcomes score threshold would be to adopt more selective admissions practices, not to focus its efforts on improving the learning environment. This statement is not meant to deny that universities can and do increase student core educational outcomes; clearly they do. It is meant to suggest that the greatest gain in achieving the preset core educational outcomes threshold with the expenditure of the fewest resources would come from increasing selectivity in admissions, not from altering the educational process to produce better core outcomes. Evidence of this can be garnered from examining the strategic efforts initiated by universities that set out to improve their standings in the *U.S. News and World Report College Rankings*. Among their strategic actions almost always will be efforts to increase in admissions selectivity.[20] Seldom will their strategic efforts focus on improving student learning.

Providing even greater incentives for universities to become more selective in admissions is not responsive to this country's need to increase the proportion of high school graduates who enter and complete higher education. The widely reported fall of the U.S. to eleventh in the proportion of 25 to 34 year olds with tertiary education

[20] For a thorough discussion of the impact of *U.S. News and World Report College Rankings* see *College Rankings Reformed: The Case for a New Order in Higher Education,* Kevin Carey, Education Sector, September 2006.

attainment[21] is adequate justification for this position. Individual public universities legitimately may pursue missions that involve increased selectivity, as part of their state's strategy to ensure that its array of institutions provides appropriate post secondary educational venues for each segment of the population. But inclusion in VSA of an element that served to propel all universities toward becoming more selective in admissions would ill serve the cause of widening access. Thus there is social gain in using value-added core educational outcomes in VSA.

Measurement of value-added to core educational outcomes reflects change in ability while the student is at the university.[22] The only obvious way a university can increase value-added is by improving the learning environment. The use of value-added core educational outcomes scores provides significant incentive for a university to focus on improvement of the learning environment. A given university or system, of course, can establish the minimum level of core educational outcomes it expect graduates to achieve but, if it does so, that should be done in conjunction with establishment of a parallel expectation that it will also produce learning gains in its students and will measure them.

4- How do we use the results of value-added core educational outcomes testing for curricular improvement?

Contrast the results of value-added core outcomes testing results with the typical university curriculum and you will understand the basis of the concern that prompts this question. The results of core-outcomes testing are at the level of the university while most decisions about the curriculum are made by the faculty member who teaches a course or by assemblies of faculty at the departmental and college level who approve course descriptions and degree requirements. How does one use the results of outcomes testing to produce change when its results may not be perceived as relevant within the spheres of influences of those who control the curriculum?

This is an interesting question but it is ultimately not the important one. The important question is, "How can universities create a focus on the whole of the curriculum, indeed, on the whole of the university experience, such that it can be altered if necessary to ensure that graduates possess those high-level skills and abilities that a truly well-educated individual possesses? The discussion above on our tendency to

[21] OECD Fact book 2008: Economic, Environmental and Social Statistics - ISBN 92-64-02946-X - © OECD 2008

[22] The words "while the student is at the university" were chosen carefully. In truth, we do not know for certain whether maturation leads to the value-added or whether the university experience itself leads to it. We do know from CLA administrations that there is variation in value-added across universities tested that is fairly consistent over time. Thus, it appears that variations in the experiences offered at different universities are associated with variation in value-added. We cannot rule out the hypothesis that at least some of the apparent value-added comes from maturation as there has of yet been no appropriate control group of young people who did not attend college tested. Note that it will be very difficult to construct such a control group because the decision to go to college or not to go indicates that two sets of young people differ systematically and casts doubt on the notion that one group is statistically able to control for the other.

"stove- pipe" the curricular experience and its assessment is relevant here. The whole purpose of value-added core educational outcomes assessment is to ensure that our focus is on the properties every graduate has rather than the properties that each of our stove-piped components of the curriculum have.

Thus concerns that core outcomes assessment do not align with goals of individual courses or segments of the curriculum are largely irrelevant. Such concerns remind me of the familiar story of the inebriated man who looked for his car keys under the street light rather than where he lost them because he could see better there. No, the results of value-added outcomes tests merely reveal whether the university's students are below (or at or above) the level of value-added they statistically are expected to be. The university and its faculty then have to engage in the research to find out why and to determine what has to be changed to move them to the expected level of value-added. Looking for the source of the problem only in established courses or groups of courses may be convenient, but it may also be a waste of time.

The story of Kalamazoo College's search for meaning in CLA scores is instructive. Kalamazoo's research was conducted by four faculty members who dug into the College's CLA scores to try to find why a subgroup of students, its science majors, did not score "above expected" as did the balance of the campus[23]. After correlating the CLA score data with transcript, NSSE and personal interview data they ultimately concluded that a disproportionate number of science majors tended to concentrate on science to the exclusion of just about everything else in the learning environment. These students merely "went through the paces" in their study abroad experiences and language classes, biding their time until they could get back to their science. Of course, on the job after graduation these students are going to lack some critical skills that their colleagues have and may well be disadvantaged in progressing in employment. Thus the fault was not so much with the curriculum as with finding a way to cause the science students genuinely to engage with it. Having gotten to the source of the problem, Kalamazoo's faculty is now much better positioned to tailor a solution.

We don't really know much about causes of variation in value-added core educational outcomes. As career faculty we tend to think that it is variation in the curriculum. This may be the result of our myopia. Some of the research into NSSE and CLA is finding that certain forms of engagement relate to value-added.[24] Living arrangements, work on campus, freshmen interest groups, library resources, undergraduate research opportunities, Greek life, study abroad opportunities, service learning, etc., might prove to be the difference between above expected value- added and lower than expected. Since many of these places were not "under the lamp post," past searches for causes of variation in core educational outcomes often have not even investigated them.

---

[23]For an excellent example of such an investigation see *MULTIPLE DRAFTS OF A COLLEGE'S NARRATIVE*, by Paul Sotherland, Anne Dueweke, Kiran Cunningham, and Bob Grossman PEER REVIEW, AAC&U, SPRING 2007|

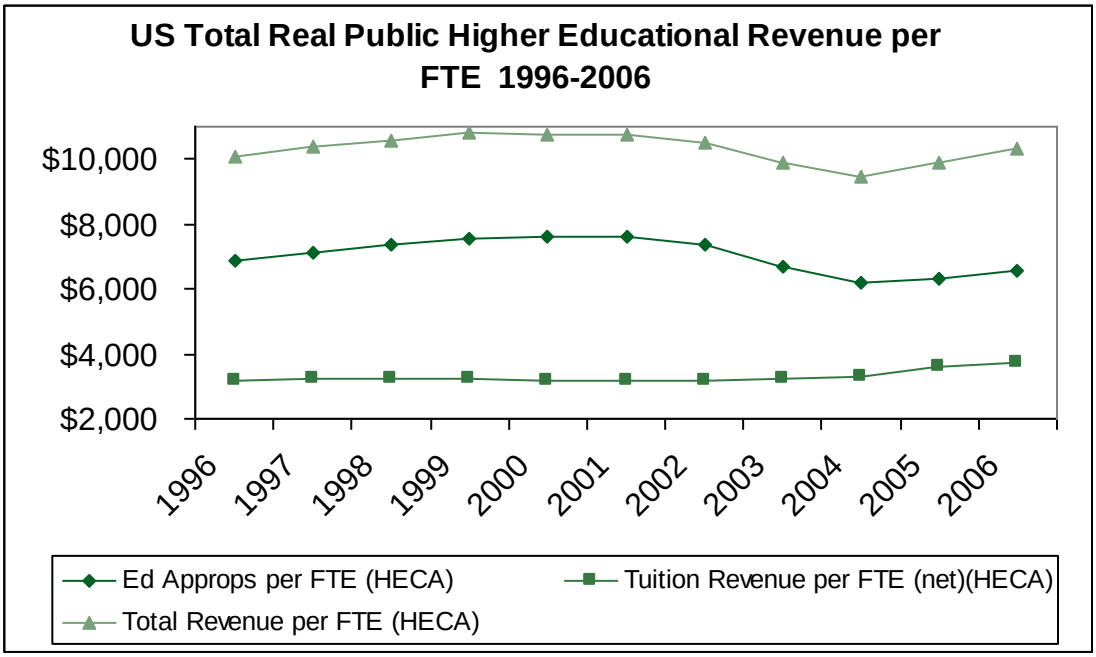
[24] See, for example, Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education, 47* (1), 1-32.

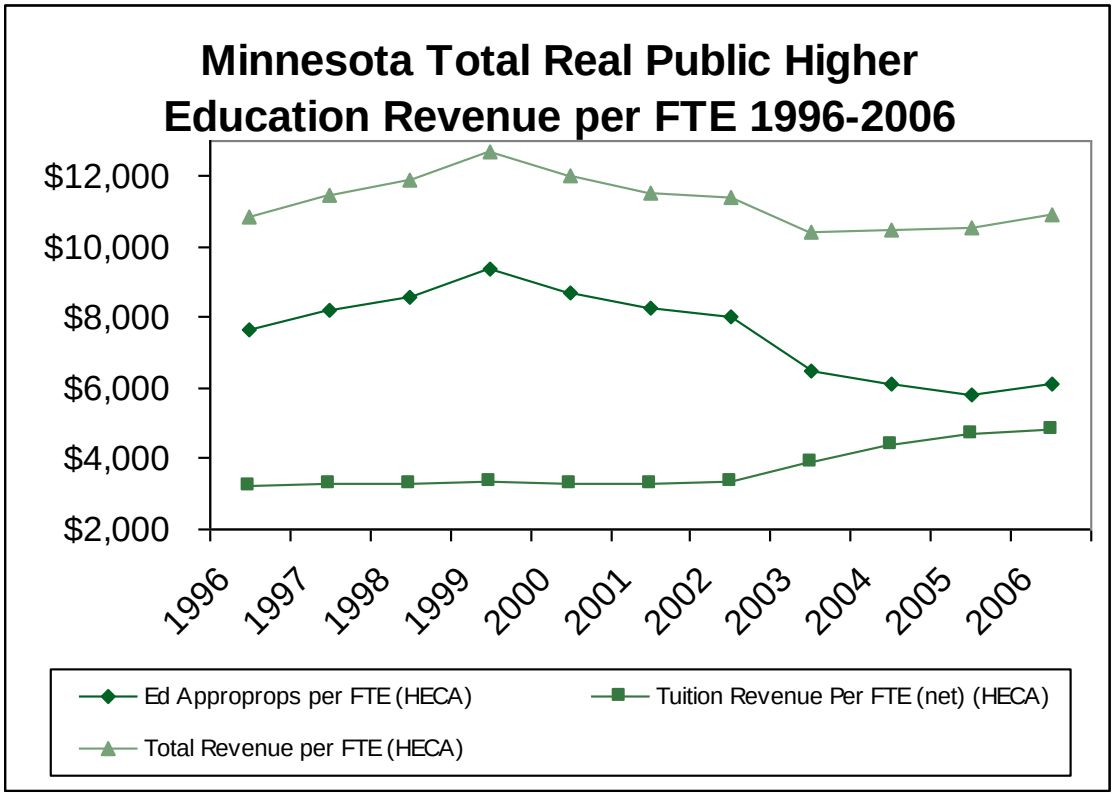
Thus, instead of bemoaning the fact that outcomes test results do not directly point us to elements within the curriculum that need changing, we ought to celebrate the opportunity those results provide us to delve into the potential efficacy of every part of the learning environment that constitutes a great university. Perhaps having such results will help us understand which of the expensive or inexpensive campus peculiarities are worthwhile as we come to understand their impact (or lack of impact) on core educational outcomes. Perhaps findings from such explorations can take some of the arbitrariness out of campus funding decisions.

As an aside, a common concern of accreditors is that the general education and major assessment results that campuses so copiously generate often have had no discernable effects on the curriculum. While we make much of the claim that "assessment belongs to the faculty" that phrase is apparently not evidenced in the activities, or hearts and minds of the faculty. It is my hope that the discovery that students on one's campus score far above expected in value-added to core educational outcomes will be a source of pride to the entire campus community and motivate efforts to understand its origin and preserve the extraordinary levels of student learning. It is easy to do harm to an ecology when one does not understand the role each piece of it plays. Similarly, the campus that scores far below expected value-added probably will be powerfully motivated to discover how to create a better learning environment.

5- **How do universities afford yet more testing**? This question is not meant to be rhetorical. The real resources of public universities on a per student basis have been roughly constant for over 20 years. Below is a recounting of that data from the State Higher Education Officers Organization, both for the nation as a whole and for Minnesota. The pattern in the two graphs is similar. Total revenue per FTE student remains about at the level of 10 years ago, state appropriations are down and net tuition revenue is up. In Minnesota, as is the case in most of the other forty-nine states, one adds a new cost element (like core educational outcomes testing) to the education budget by economizing on other expenditures. Growth is by substitution, not by the less painful route of addition.

While the public and legislators don't seem to be aware of this reality, public academic administrators have to live with it every day. While the data set I present below is for the last decade, taking it back two decades does not change these patterns. Will the patterns change in the next decade? I'm not clairvoyant, but what is happening in most state capitols as the increasingly weak economy negatively affects state revenues is a continuation of cyclical variation of state appropriations. Real per student appropriations once again are falling in most states. In many states, tuition increases have been substantial. I suspect total real revenue per FTE will not increase, at least this year.





Can states be persuaded to pay for core educational outcomes testing? Perhaps, but remember that the data shown above includes all those special appropriations for things like computers, fuel adjustments and catch-up salary increases that have occurred over the past decade. Those appropriations merely

reshuffled the deck chairs; they did not elevate the deck. My point is that public higher education is often successful with requests for ear-marked appropriations but these ear-marks are generally offset, leaving total appropriations unchanged. Thus experienced (read "jaded") administrators are not counting on genuine "additional" appropriations for this purpose.

Thus the imperative is to find accurate and effective low-cost methods to measure value-added core educational outcomes. Fortunately, each of the three tests used in VSA is relatively inexpensive; the direct cost estimates for administration of the tests on a sample basis are in the accompanying text box. The key word in the preceding sentence is "sample." For a university with 40,000 students, testing all of its freshmen and all of its seniors, direct costs would range from $300,000 to $650,000 depending upon the test chosen. There are, of course, additional logistical costs associated with administering exams and, perhaps, educational costs if one has to suspend classes for a day so that nearly all students can take one of these tests.

CAAP (200 students each in fall and spring):	$10,200
CLA (100 students each in fall and spring):	$6,500
MAPP (200 students each in fall and spring):	$6,200

Sampling is necessary to keep the costs of testing modest. Samples of a reasonable size can provide accuracy equivalent to that of measuring the entire population. Each of the three test makers recommends sampling; CLA recommends a sample of 100 students in the freshmen and senior classes and CAAP and MAPP recommend testing 200. Each testing firm recommends that the sample drawn be a reasonable reflection of the campus community.

But use of a sample has its own costs. In this case, the costs are those of inducing the students in the sample to show up to take the test and to take the test seriously. These tests are fundamentally unlike the Graduate Record Exam or licensure exams because there is no personal consequence to the test taker for turning in a good or bad performance. Thus universities may have resorted to artificial inducements to get students to take the test and to perform at their ability level. Such inducements are most frequently of a monetary form, but sometimes take the form of gift certificates, priority in registration or other queues, gifts from the bookstore, and appeals to the students' loyalty to the alma mater. While the test makers contend that biases arising from the use of student participation inducements have little material impact on scores that cannot statistically be controlled for, members of the academy tend to have suspicions of data generated by subjects who participate because of pecuniary inducements.

Most (including the VSA participants) use cross-sectional test administration schemes rather than longitudinal schemes in an effort to minimize the costs of using a sample. This introduces interpretative considerations. Using a cross-sectional

scheme to measure value-added by the educational process assumes that the experience freshmen will have essentially will replicate the educational experience the seniors had, i.e., that the educational environment of the university will be unchanged for a seven-year period. Thus careful documentation of changes in the educational environment over time is essential if one is to fully understand factors that contributed to value- added.

On the other hand, longitudinal testing assumes that the cohort of seniors that survives from the entering freshman class is randomly drawn from that original freshman cohort. While one can control for differences in cohorts based on available demographic and admissions test scores (ACT/SAT), the suspicion that the surviving seniors differ on other important dimensions from the freshman cohort is hard to rebut. In addition, a longitudinal sampling plan is much more costly to implement than a cross-sectional plan. The additional costs arise from the need to over-sample the freshmen class to ensure that enough freshmen survive to the senior year to provide an adequate sample of seniors, tracking the students over four years to ensure that they can be included in the senior sample, and inducing the student who may not have great memories of having taken a test when a freshman to take the same test again as a senior. Some refer to longitudinal testing as the "gold standard" method of value-added testing, but practical experience of those who have used the longitudinal method suggests that the method may be of a decidedly less attractive hue. It is very, very difficult to track and induce repeated attention to testing from a continuing cohort of students over a four year period of time; this makes longitudinal testing significantly more expensive than cross-sectional testing.

Whether cross-sectional or longitudinal sampling schemes are used, it is essential to control for differences in the academic quality of students in the freshman and senior samples (as measured by their entering test scores). Doing so provides the best opportunity for measuring real value-added attributable to the educational program rather than "value-added" that results from attrition of lower ability students from the sample. It also makes it impossible for an ethically challenged university to artificially create value-added by testing a less academically able sample of freshmen and comparing their scores with those of a more academically able cohort of seniors. Note that use of the uniform method of measuring value-added that was described above (the method to be used by VSA participants) controls for freshmen/senior differences in initial ability by creating expected scores as the appropriate indices for comparisons.

Over the next several years universities participating in VSA will learn much from their trial administration of these measures about how to minimize the potentially data-distorting properties of sampling. My best guess is that many universities ultimately will decide to embed their core educational outcomes testing in freshmen and senior courses that enroll a nearly random set of students from the two classes. The outcomes of the exams will figure somehow into the course requirements in order to ensure that the student understands that the results have

consequences. Most sampling and motivational problems are satisfactorily dealt with by embedding.

Most universities will find that administration of core educational outcomes tests on an annual basis yields little useful data and decide to reduce the frequency of administration. After an initial few years during which the university may benefit from annual administration (in Deming's terms, getting the process into "control"), reducing the frequency of administration to every third year is generally appropriate. Doing so reduces the cost of the testing program by two-thirds. After that, only major changes in curricular policy or admissions policy would warrant more frequent administration.

On the other hand, I also suspect that most universities will want to increase samples beyond the bare minimum size needed to understand the value added to the learning process for the average student. They will want to know the degree to which value-added varies by student demographic characteristic, by their living arrangements, work experiences and student activity participation patterns. Variations in value-added for all of these subcategories cannot be derived from a minimal sample, but modest enlargement of the sample can enable meaningful analysis by subcategory. Preliminary research results that I have seen suggest that students from some demographic groups exhibit less value-added than those from other demographic groups even after statistically controlling for entering student test scores. Should these study results hold up under further academic scrutiny, universities will want to develop interventions to eliminate or lessen the value-added differences. They, of course, can do so only if they know what those differences are, so sampling that permits identification of differences in value- added core educational outcomes by subgroups generally will be necessary.

Remember that my context in this discussion is the use of learning outcome testing to measure institution-specific value-added to core educational outcomes. Another use of core educational outcomes measurement would be to create a credential that a college graduate could use to demonstrate readiness to work. Testing for such an objective would either have to be required of all college graduates or else be an option for those graduates who wished to receive the credential. Groups like the Partnership for 21st Century skills have developed inventories of skills that high school, community college and college graduates should have and advocate developing and using such tests. Their skill inventories include high level core educational outcomes like critical thinking, but also include many other specific skills. Tests like the CLA, CAAP and MAPP would have to be augmented with a battery of additional tests to produce results useful to employers in evaluating the desirability of employing a specific graduate. Such testing would be extremely expensive, both because of its breadth and because of the precision of measurement needed if it is to be both fair to the student and helpful to the employer. It is sufficient to distinguish testing for accountability and university improvement as discussed here from testing for the purpose of creating ability-certifying credentials. The former is affordable within existing university budgets; the latter probably is not.

Conclusion

In conclusion, I refer again to the graphs on U.S. and Minnesota higher education real revenue per FTE student. Both show declines in real per student appropriations over time. Minnesota's decline is substantial, from a peak of $9,345 in 1999 to $6,100 in 2006.

Why is the average student's higher education worth 35% less to this state than it was seven years ago? The most frequently ventured hypotheses likely will range from "firm obligations of the state (e.g., prisons, the state's portion of federal entitlements, etc.) have grown so rapidly that the state cannot maintain the fiscal support to higher education that it once afforded" to "the private return to education has grown to be so large that the State feels less obligation to pay as large a portion of total educational costs than it did historically."

I'm neither a Minnesotan, nor an expert on Minnesota public finance and have not studied historical trends enough to venture a hypothesis specific to this state. What I do want to venture is the notion that this country's confidence in public institutions, including public higher education, has declined in recent years and the widespread declines in appropriations are a consequence. Tax decreases are easy sells. Criticism of public institutions has launched many a successful campaign for public office.

Recently voiced skepticism about the value of a college education from Charles Miller, the chair of the Commission on the Future of Higher Education, is pointed:

> *There are powerful reasons to question the conclusions . . . that "education pays" and significant evidence exists that an opposite set of conclusions could be reached. . . . using assumptions more in line with current realities might reach the shocking conclusions that American higher education today has gotten too expensive for what it produces; that it has become too costly for the typical student…..that education (a college degree) does not pay!*[25]

Few outside of higher education made the effort to rebut the Mr. Miller's contention. I suspect that the lack of a stirring defense by societal leaders is because his words ring somewhat true to many of them. In an environment in which such allegations can go essentially unchallenged, it is not surprising that legislators in almost every state are reducing their commitment to higher education.

Silence is not our friend in this environment; the effective voice that we need will not be filled with rhetoric that repeats accomplishments of the past backed by recitals of

[25] Personal letter form Charles Miller to Gaston Caperton, President, College Board, dated April 2, 2008 as made electronically available by Inside Higher Education on April 7, 2008.

the abstract virtues of higher education. Our detractors allege that we are unproductive, wasteful and that our students benefit less than we have claimed. We must respond with data. Data for higher education in general will not be adequate; data that relates to specific universities is required as funding is not provided by the public to higher education in general, but to specific institutions.

Data on value added to core educational outcomes responds to the public's call for more and better information. Data alone will not be sufficient, but it is a part of the response required. Accompanied with specific data on the post-graduation success of each institution's graduates, the objectively measured impact of the institution's research and tech transfer on regional economies, and fully transparent data on the efficiency and integrity of the fiscal management of each university can be put forward as a package that will serve as a powerful antidote to public skepticism. If it accomplishes nothing else, generating and publishing transparent, comparable and meaningful data will serve to diminish the volume of those who believe we are hiding something.

We can no longer get by with "trust us." Measurement is required.

Appendix I

VOLUNTARY SYSTEM OF ACCOUNTABILITY (VSA) PROJECT COMMON QUESTIONS FOR TEST VENDORS

Please answer the questions as directly and as thoroughly as possible. Do not refer the reader to outside publications or sources. Answers to items I-A and II-A are particularly importance to the selection decision as the VSA project is committed to the use of gain scores or value added scores

I. Gain Scores and Value Added

A. Regardless of your past practices, will you develop a method of using your test to generate and report "gain scores" or "value-added" scores at the institutional level?

B. If not, why not?

C. If so, what method will you use to calculate them? Please explain your methodology in detail.

II. Reporting Conventions

A. Will you report institution-level scores controlled for the academic ability of each university's students 1) as measured by ACT/SAT and 2) as measured by high school GPA combined with ACT/SAT?

B. Can you generate score reports that would indicate “above expected level of performance,” “at expected level of performance,” and “below expected level of performance?”

C. If you cannot or will not, please explain why and please describe the form in which you can report scores.

D. Would you make the results available in an electronic format that would be suitable to merge with institutional student records? What type of file would be available?

E. Do you offer institutions the opportunity to compare their results with the scores of other institutions either individually and/or within categories/groups? If group comparisons are available, does the institution have the ability select their own “peer” group?

III. Miscellaneous Relationship with other Measures

A. Are there linkages or relationships between your test and any standardized placement test, e.g. a test used to determine what initial math or English course a freshman should take, such that the placement test could serve as a control for the entering ability of students?

B. How is relationship between ACT/SAT scores and the scores on your test calibrated? How stable is the calibration between ACT/SAT and the scores on your test? Do you anticipate that the relationship will need to be recalibrated? If so, how often?

C. Are there known correlations between student performance on your test and ACT sub scores (English, math, reading, science) or SAT sub scores (critical reading, math, writing)? If so, what are the correlations?

D. Would you be willing to develop a guide that explicitly delineates the correlations or equivalent scores between your test and other tests?

IV. Availability of Individual Student Scores

A. Will you report individual student scores to each institution?

B. Will you provide individual scores or some other form of performance report to individual students? Please elaborate.

V. Constructs Tested – The AASCU/NASULGC Voluntary Accountability System has identified three constructs to be measured by selected instruments: critical thinking, analytic reasoning and written communications.

A. Regardless of how you report scores now, would you be willing to develop and report scores for each of these three constructs: critical thinking, analytical reasoning and written communications?

B. If so, how would you generate those scores from your test?

C. If not, please explain why you would not do so?

D. What other scores will you report?

VI. Sampling Protocol

A. What is your recommended sample selection process for an institution that wishes to measure gain scores or value added for undergraduate students across the entire university?

B. What is the minimum sample size you recommend for this purpose?
C. What is your recommended sample selection process for an institution that wishes to measure overall gain scores or value added for undergraduates at the institution level as well as subgroups such as academic programs or student demographic groups?
D. What is the minimum sample size you recommend for this purpose and how does it vary by the number of subgroups.

VII. Student Motivation

A. What are your recommended strategies for motivating student to participate in the assessment?
B. What are your recommended strategies for motivating students to perform at their ability level on the test?
C. Do you have any method to identify students who are "blowing off" the test or who are experiencing "test fatigue?"
D. What do you currently provide or plan to provide as a test maker to improve student motivation? (e.g., certificates of achievement, etc.).

VIII. Scoring Range

A. Is there a maximum score on each section of the test?
B. What proportion of four year college students achieves 95% of the maximum score on each section of the test? What percentage achieves the maximum score?
C. Is the test sensitive enough to find increases in learning for both open admissions and highly selective universities?
D. What proportion of those tested in highly selective universities achieves 95% of the maximum score on each section of the test?
E. What is the relationship between gain scores/value added scores measured by your test at the institutional level and 1) entering ACT/SAT institutional-level scores, and 2) a combination of ACT/SAT and high school GPA levels?

IX. Transfer/Non-traditional Students

A. Do you have a recommended method for using your test to estimate institution-wide gain scores or value added score for transfer and/or non-traditional students? If so, please describe that method in detail.
B. For students who do not have ACT/SAT scores do you use some other measure(s) as a proxy? If so, what measure(s)?
C. Do you have a recommended strategy for selecting a sample for the purpose of developing an institution-wide gain score or value added score for transfer and/or non-traditional students? If so, please describe in detail.
D. Comment on the utility of comparing the learning outcome scores of native seniors institution wide with those of transfer and/or non-traditional seniors institution wide.

Appendix II

FIPSE Construct Validity Study of Core Educational Outcomes Tests

The Education Testing Service (ETS), the American College Testing Program, Inc. (ACT) and the Council of Aid to Education (CAE) are collaborating to examine the construct validity of various critical thinking and writing measures (i.e., MAPP, CAAP and CLA) that are options for use in the Voluntary System of Accountability program. In the context of this program, researchers will address the research questions below by giving different combinations of these measures to approximately 1,200 students (freshmen and seniors) at 13 colleges and universities across the nation in fall 2008.

Key Project Personnel

Dr. David Shulenburger, Vice President for Academic Affairs, NASULGC, will oversee this portion of the project and he will report to Dr. Terry Rhodes of AAC&U. Dr. Stephen Klein will serve as principal investigator. He will be assisted by Alex Nemeth (project manager), and Dr. Roger Bolus who will be in charge of data management and statistical programming. The ACT portion of the project will be led by Dr. James Sconing. The ETS portion of the project will be led by Dr. Brent Bridgeman.

Research Questions

1. Do different measures that are designed to assess the same construct (such as critical thinking) correlate higher with each other than they do with tests that are designed to assess other constructs (such as reading)?

2. 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 higher with each other than they do with scores on tests that use the same response mode but assess different constructs? In other words, how much of a student's score on a test is a function of response mode and how much is it due to the student's mastery of the underlying construct being assessed?

3. After controlling on sample characteristics, do different measures of the same construct have similar differences in mean scores (effect sizes) between freshmen and seniors?

4. What is the reliability of student and school level scores of different measures of writing and critical thinking ability? And, how many students are needed to obtain reasonably reliable school level scores?

5. Do different measures of presumably the same construct lead to the same versus different conclusions regarding whether a school's seniors performed at, above, or below what would be

expected where the expected level is based on (a) their academic ability level at the time they entered college (as measured by ACT/SAT scores) and (b) the typical relationship between the average ability at a school and its students' average scores on the project's tests.

Study Timeline

2008	
Aug. 1 – Sept. 20	Proctor training and delivery of materials
Aug. 20 – Sept. 10	Testing at pilot school
Sept. 10 – Sept. 20	Window for correction of unanticipated issues at pilot school
Sept. 20 – Oct. 31	Testing at all schools
Oct. 31 – Nov. 15	Testing window cushion (if necessary)
Nov. 1 – Nov. 20	Registrar data collection
Dec. 20	Scoring and registrar data processing complete
2009	
Feb. 15	Common data file compiled
Feb. 15 – April 1	Data analysis
May 1	Report drafted
June 1	Final report complete

As of June 6, 2008:
MIT is serving as a partial pilot school to test out procedures with a few students. The steering committee has been meeting/holding conference calls regularly. Most of the work lately has been taken on by the "Special Operations" team, which has been developing student fliers, student tracking sheets, study ID cards, counterbalancing guidance, session scheduling/tracking sheets, and a test(s) administration manual.

Participating Schools

1. Alabama A & M University
2. Arizona State University at the Tempe Campus
3. Boise State University
4. California State University, Northridge
5. Florida State University
6. Massachusetts Institute of Technology
7. University of Colorado at Denver
8. University of Michigan-Ann Arbor
9. University of Minnesota-Twin Cities
10. University of Texas at El Paso
11. University of Vermont
12. University of Wisconsin-Stout
13. Trinity College (Hartford)

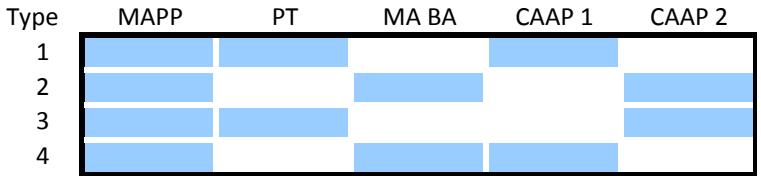
Test and Incentive Plan

	MAPP	PT	MA BA	CAAP 1	CAAP 2
MAPP		300	300	300	300
PT				300	300
MA BA				300	300

Each shaded rectangle is a cell representing students taking two of the five tests. We need 300 students in each of the 8 cells above for the correlations. Students will take three tests: the MAPP, one CLA (either PT or MA/BA) and one CAAP (either CAAP 1 or CAAP 2), each on a separate day. Students will be randomly assigned to either subtest within CLA or CAAP measures and guidance to schools will attend to counterbalancing the order of tests.

Students will receive a $150 Amazon.com gift card for three tests as an all or nothing payment. The gift cards would be distributed by CAE via email based on weekly reports from test sites that will include lists of students who have completed all three tests. Therefore, they should be received by students no more than two weeks after completion of the third test.

To estimate the number of unduplicated students, we categorize them into 4 types as follows:



We end up with more completed MAPP tests but a simpler plan: all students take three tests. We need these four types to fill the 8 cells in the final test matrix. As such, the unduplicated number of students across the study is 4 student types X 300 = 1,200 (roughly 92 at each school).

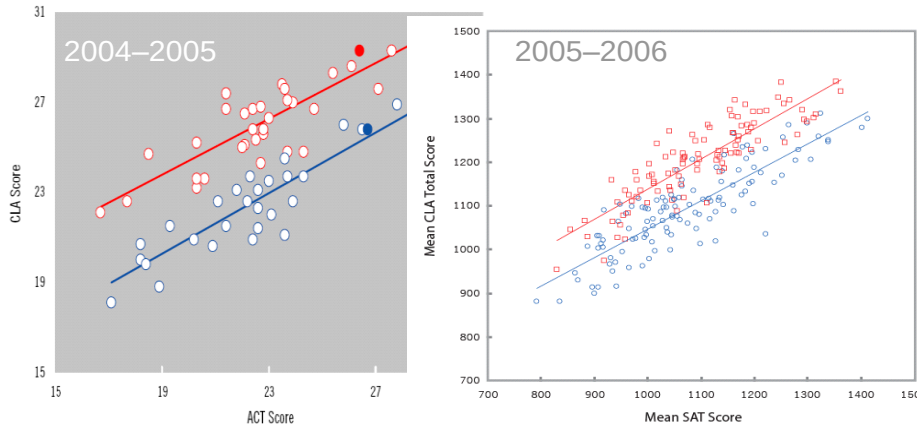
Campus Teams

Campuses receive a $4,500 stipend, a portion of which will be tied to successful implementation..

- Campus facilitator: The campus facilitator will (1) Recruit, supervise and assist Field Coordinator; (2) Ensure adequacy of testing facilities; and (3) Support resource allocations. Must be a full-time university staff member.
- Field coordinator: The field coordinator will be responsible for (1) Recruitment of students and meeting testing goals; (2) Proctoring; (3) Test security; (4) Weekly updates to CAE; (5) Returning test materials to the testing agencies. May be a graduate student, but may not be an undergraduate student.
- Examination Proctors: The proctors will be responsible for administering the exams after undergoing web-based training. May be graduate students. May not be undergraduate students.
- Registrar: The students' college GPA and admission test scores will be obtained from their registrar's office. The registrar may also be asked to confirm student-entered information, such as class standing.

Appendix III

Value-added models consistent over time



CLA Total Score Regression Equations for Freshmen and Seniors

	Freshmen		Seniors		Difference	
	Slope	R-Square	Slope	R-Square	Slope	R-Square
2005-2006	0.65	0.74	0.69	0.76	0.04	0.02
2004-2005	0.66	0.80	0.62	0.75	-0.04	-0.06
Difference	-0.01	-0.06	0.07	0.01		

9-33 02/06

Figure 1: Relationship between CLA Performance and Incoming Academic Ability