National Institute for Learning Outcomes Assessment

September 2012

The Seven Red Herrings About Standardized Assessments in Higher Education

Roger Benjamin

Foreword by Peter Ewell

Commentaries:

Margaret A. Miller

Terrel L. Rhodes

Trudy W. Banta and Gary R. Pike

Gordon Davies

Occasional Paper #15

learningoutcomesassessment.org

About the Authors

Roger Benjamin has been President of CAE (Council for Aid to Education) since 2005. He was a research scientist at RAND from 1990 to 2005 (director of RAND Education, 1994-1999). Previous to his appointment to RAND, he was professor of political science at the University of Minnesota, 1966 to 1983 and associate dean and executive officer, College of Liberal Arts, University of Minnesota, 1980 to 1983, vice chancellor for academic affairs and provost at the University of Pittsburgh, 1983 to 1986, and vice president for academic affairs and provost, University of Minnesota, 1986 to 1988, and professor of political science, 1988 to 1990. He is the author or co-author of numerous books, monographs and articles in political economy and public policy, including governance, strategic planning, and assessment in education policy. His latest book is *The New Limits of Education Policy: Avoiding A Tragedy of the Commons*. London: Edward Elgar, 2012. He directs a program implementing performance assessment throughout the K-16 education system in the United States and beyond.

Margaret A. Miller is a professor in the Center for the Study of Higher Education at the Curry School of Education at the University of Virginia, president emerita of the American Association for Higher Education, and editor-in- chief of *Change* magazine.

Terrel L. Rhodes is the Vice-President for the Office of Quality Curriculum and Assessment at the Association of American Colleges and Universities (AAC&U). He is also director of the annual AAC&U General Education Institute.

Trudy W. Banta is Professor of Higher Education and Senior Advisor to the Chancellor for Academic Planning and Evaluation at Indiana University–Purdue University Indianapolis. She is the founding editor of Assessment Update, a bimonthly periodical published since 1989 by Jossey-Bass.

Gary R. Pike is the Executive Director of Information Management and Institutional Research at Indiana University Purdue University Indianapolis, and Professor of Higher Education.

Gordon Davies served as the Director of the State Council of Higher Education for Virginia and as President of the Kentucky Council on Postsecondary Education. He currently is a senior advisor to a Lumina Foundation project.

Peter Ewell, NILOA Senior Scholar, is the Vice President at the National Center for Higher Education Management Systems (NCHEMS), a research and development center founded to improve the management effectiveness of colleges and universities.

Contents

Abstract

This occasional paper by Roger Benjamin outlines the merit and role of standardized tests for assessment in higher education by addressing familiar arguments against standardized assessments that have confused participants on each side of the debate about the need for and the possibility of new benchmarks on student learning outcomes. Benjamin argues that the key seven assertions, or red herrings, need to be set aside in order to achieve progress toward the goal of continuous improvement in student learning outcomes. In his foreword, Peter Ewell sets the context for Benjamin's position. Four commentaries by higher education thought leaders knowledgable about assessment examine further the promise and pitfalls of using standardized tests to measure and enhance student learning.

Foreword

Debates about the merits of standardized testing as a measure of college and university effectiveness have been around since the assessment movement began. Indeed, I recall attending in 1986 a spirited meeting on this topic convened by the Assessment Forum of the American Association for Higher Education (AAHE) that resulted in a publication cautioning faculty against wholesale reliance on standardized tests as a vehicle for assessment.[1] So, by now, the focus of this collection of essays, centered on Roger Benjamin's defense of standardized testing in the form of "seven red herrings," is familiar.

Despite considerable divergence of opinion among them, the authors of these essays are in basic agreement about several points. Probably the most important of these is that standardized testing should not be the only instrument for measuring student learning outcomes that institutions and policymakers employ. Similarly, all of the authors agree that any measurement instrument should be used both to demonstrate current levels of achievement and to guide improvements in teaching and learning. The authors also express a good deal of support for the notion that external benchmarks of achievement are useful, although Rhodes as well as Banta and Pike caution that the most useful points of comparison will not involve comparisons among institutions. The two areas of substantial disagreement among these authors, however, are classic debates.

The first of these debates concerns the extent to which commercially available examinations such as the Collegiate Learning Assessment (CLA) or the ACT Collegiate Assessment of Academic Proficiency (CAAP) are really "standardized" in the sense that they yield valid and reliable results that can be compared across settings. Benjamin advances the conventional argument that the major virtue of such tests is that they are carefully and deliberately designed to generate comparable data. Banta and Pike argue, however, that the artificial settings divorced from actual classrooms in standardized testing, as well as uncontrolled and unknown differences in student motivation to perform well on tests that don't count, will always doom comparisons between different populations. At the same time, Rhodes argues that faculty raters can be just as "standard" as multiple choice tests when they evaluate student work if they are properly trained and if they use well-designed rubrics. Miller disagrees, pointing out that professional test makers are far better prepared to create questions and examination conditions than faculty who have had next to no training in these areas.

The second area of substantial disagreement among these authors is the extent to which "generic" competencies such as critical thinking or communication can even be measured independent of discipline content. Benjamin maintains that abilities like how to locate and evaluate information have become far more important than simply possessing content knowledge, as the volume of the latter has grown enormously and has become better documented. He also claims that student performance on standardized generic skills examinations is unaffected by what a student currently knows or is studying, and he backs this up with studies performed on the CLA. Miller agrees, pointing out that employers value such skills and arguing that not to teach them would be "a version of educational malpractice." Banta and Pike provide evidence to the contrary, in part drawn from the Council of Independent College's "Value Added" project and from an evaluation of the Voluntary System for Accountability (VSA) indicating that a test taker's major field can have a marked impact on CLA scores; at one college, including or not including nursing students in successive years caused noticeable fluctuations in institution-level value-added scores. Interestingly, the Educational Testing Service (ETS) designed

[1] Heffernan, J. M., Hutchings, P., & Marchese, T. J. (1988). *Standardized tests and the purposes of assessment*. Washington, DC: AAHE Assessment Forum, American Association for Higher Education.

Foreword continued

the prototype of the CLA—the Tasks for Critical Thinking—well aware that discipline context would matter in task performance. Accordingly, ETS required test takers to undertake tasks set in the sciences, social sciences, and humanities to neutralize this effect. The current CLA does not follow this approach, and sets performance tasks in "real world" contexts instead. Moreover, Banta and Pike, based on their experiences many years ago at the University of Tennessee, Knoxville, take Benjamin to task for using value added—a measurement approach that the VSA used for many years but that Banta and Pike believe is badly flawed.

While advancing his case on the merits of standardized tests, Benjamin makes two broader points about assessment that the other authors also engage. First, he advances "institutional inertia" as the primary explanation for the reluctance of colleges and universities to embrace assessment wholeheartedly despite it being in their interest to do so. This point is the central theme of the essay by Davies, who asserts (correctly, I believe) that the values of higher education institutions are driven by an essentially conservative reward system based largely on prestige. In such an environment, superior learning outcomes do not count for much compared to large endowments, selective admissions, and well-recognized faculty research agendas. As a result, Davies concludes, colleges and universities engage in assessment only when they are forced to do so by external authorities—state coordinating and governing boards or regional accreditors.

A second point that Benjamin makes, almost in passing, is that testing organizations should not report scores publicly themselves but should supply results directly to institutions which, in turn, should be encouraged to report their scores themselves—presumably through a mechanism like the VSA. Although his position on this matter is not prominent in Benjamin's argument, it provoked strong rejoinders from both Davies and Miller. Because institutions are likely to report results selectively to suppress bad news and to emphasize success, this position is a "dead end," in Davies's view—a view corroborated by institutions' frequently inappropriate use of their scores on the National Survey of Student Engagement (NSSE). Taking a point from Atul Gawande about driving improvement in medical practice by posting outcomes publicly, Miller believes that similar action in higher education will drive colleges and universities to improve, if only to avoid embarrassment, and, further, that public disclosure will make superior performers more visible so that others can learn from them. It is perhaps no coincidence that both Davies and Miller have more experience at state boards than in running institutions.

Taken collectively, the arguments in these essays about the use and appropriateness of standardized tests in higher education demonstrate why this subject won't go away. In short, there are at least three reasons why higher education institutions (and the faculty who inhabit them) do not like standardized tests:

- Higher education institutions and faculty do not control the contents of such instruments. Because American higher education remains fiercely faculty centered, it tends to resist approaches to assessment that are not individually tailored to the diverse and idiosyncratic content areas typically present across institutions in even the same disciplines. So even though standardized tests may, in fact, provide more precise measurements than locally designed assessments, they fall victim to faculty's innate need to maintain control over what is taught and assessed.

Foreword continued

- Faculty don't like giving money to testing organizations. Testing is big business and faculty members are suspicious of any organization that has commercial purposes. Investing resources to buy tests from outside the institution exacerbates this natural distaste. Home-grown approaches like rubrics and portfolios, while expensive in faculty time, involve reallocations of internal spending rather than a net outflow of resources.

- Faculty believe that standardized tests smack of accountability and they know that they are popular with external authorities. Legislators and state officials like evidence that is numeric and seemingly simple to grasp, and they are also already familiar with test data used for accountability in elementary and secondary education. As evidenced by several of the authors in this collection, the current provisions of No Child Left Behind are frequently cited in cautionary tales for assessment in higher education.

Each of these reasons, of course, has an opposing rationale and an associated constituency. Although the discussion has deepened over the years, as evidenced by this collection of essays, the debates are sure to continue into the future about the appropriate merit and role of standardized testing.

Peter Ewell
NILOA Senior Scholar
Vice President, NCHEMS

The Seven Red Herrings About Standardized Assessments in Higher Education

Roger Benjamin

Introduction [1]

Why are many higher education stakeholders reluctant to accept standardized assessments? By "standardized assessments" I mean assessment instruments in which the questions, the scoring procedures, and the interpretation of results are consistent and which are administered and scored in a manner allowing comparisons to be made across individuals and groups. The list of skeptical stakeholders includes faculty, administrators, boards of trustees, accrediting groups, and membership associations. In this paper, I will state the most familiar arguments against standardized assessments in higher education that have confused participants on each side of the debate about the need for and the possibility of new benchmarks on student learning outcomes.

Seven red herrings need to be discarded if we are to achieve progress toward the goal of continuous improvement in student learning outcomes ...

The higher education community faces a paradox with respect to assessment that must be resolved. Neither standardized assessment nor formative assessment[2] alone is adequate without the contribution of the other form of assessment. It will not be easy to resolve this conundrum. My intent is to encourage the development of solutions to eliminate it.

We need standardized assessments to permit faculty and administrators to signal how well they are doing in comparison with other higher education institutions. Most importantly, we need good standardized assessment instruments to encourage the development of assessment strategies that directly help faculty to improve teaching and learning in a systemic and continuous manner.[3] Standardized instruments that permit comparison are a necessary condition for progress in developing a more systematic approach to assessment in higher education. Nuance is very important here, however. While standardized tests are necessary, they are not sufficient for integrating assessment with teaching and learning at institutions. Formative assessments developed by faculty at institutions are also critically important for any assessment system to be complete. Why has this view not yet prevailed? If the responses to arguments against standardized tests are convincing, why do the same arguments expressing the same shibboleths constantly recur? Here, then, are the key seven assertions, the seven red herrings, which need to be discarded if we are to achieve progress toward the goal of continuous improvement in student learning outcomes.

[1] This paper builds upon arguments in my recent book, *The New Limits of Education Policy: Avoiding a Tragedy of the Commons*. I thank George Kuh for his comments on an earlier draft of this essay.

[2] Formative assessment is conducted to assist faculty to directly improve their teaching and learning in the classroom.

[3] I define "good standardized assessment instruments" as tests that are backed up by substantial studies corroborating their reliability and validity and that are being used by significant numbers of colleges and college students. The principal standardized tests used in U.S. postsecondary education include the Proficiency Profile (ETS), the California Critical Thinking Skills Test (CCTST [Insight Assessment, California Academic Press]), the Collegiate Assessment of Academic Proficiency (CAAP [ACT]), the Collegiate Learning Assessment (CLA [CAE]), and the Watson-Glaser Critical Thinking Appraisal (Pearson). A study by measurement scientists at ETS, ACT, and CAE (see Klein et al., 2009) found that the critical thinking measures used in the tests developed by these three organizations were significantly correlated and, therefore, that all three tests may be regarded reliable. This does not mean, however, that these three tests measure the same thing (Steedle, Kugelmass, & Nemeth, 2010).

The Seven Red Herrings

1. Because it is impossible to measure all that is important in education, it is impossible to measure anything that is important.

This logical fallacy has provided a convenient excuse to not measure any important aspect of education with assessment instruments that meet the highest canons of scientific reliability and validity—standardized assessments. The cost of using this excuse is high. If you do not benchmark progress in important dimensions of student learning, how do you know how to evaluate your institution, department, program, or faculty members? Additionally, if an institution is unable to compare itself against its competitors, how will it know how to improve its approach to teaching and learning?

If you do not benchmark progress in important dimensions of student learning, how do you know how to evaluate your institution, department, program, or faculty members?

It is possible to measure certain important aspects of student learning, often called higher order skills—critical thinking, analytical reasoning, problem solving, and written communication—while still adhering to the scientific canons noted above. These skills are considered crucial in the Knowledge Economy by most colleges and universities (see mission and general education statements), faculty, many employers, and observers (see Bok 2006; Stevens, 2010; Wagner, 2008).

2. Comparison of higher education institutions is not warranted for two reasons and, in any event, is not necessary for a third reason. These reasons follow:

a. Missions and visions of colleges and universities are so different that it makes no sense to compare them. Furthermore, research has shown no statistical differences between institutions on measures of critical thinking—the educational component measured most often.

b. Variance is much higher within institutions than between institutions, so between-institution comparison is not worth doing.

Allow me first to respond to each of these points directly and then to elaborate my position:

a. Most higher education institutions commit to improving higher order skills as a fundamental part of their compact with students. The fact that there can be at least two standard deviations between similarly situated colleges and universities, including selective colleges[4], means there is a substantial canvas of similar institutions where researchers may study best practices in teaching and learning. Critical thinking is currently measured by at least three standardized tests, so lessons can be learned from institutions doing better than expected—lessons that can then be adapted for institutions not doing as well.[5]

b. What conclusion can adherents of the within-institution approach to assessment draw from the fact that the standard deviation for any variable (thereby, its standard error and confidence interval) is larger when the student is the unit of analysis versus when a collective group, such as a college, is the unit being analyzed? For example, an individual with an SAT quantitative reasoning score of 600 is at the 85th percentile, but an an institution with a mean SAT score of 600 is probably closer to the 99th percentile. Thus the distribution of student scores is much more spread out than is the distribution of school scores. This unit-of-analysis effect—just as true for the National Survey of Student Engagement (NSSE) scales as for the CAAP and other standardized assessments of higher order skills—does not mean that such comparisons are not worthwhile.

[4] "Similarly situated colleges" are defined as colleges with student populations that are similar based on the entering competencies of the students as measured by the ACT or SAT (see Benjamin [2008]).
[5] The measures are the CAAP, the CLA, and the Proficiency Profile.

An analogy can help put this discussion in context. The range of averages among the players on a major league baseball team is greater than the mean differences in team averages. But does anyone think that the differences between the team averages do not matter? The report on the results of a school's student satisfaction questionnaire, for example, will give the percentages of students who mark "never," "sometimes," "often," or "very often" as answers to a variety of questions. But how meaningful or interpretable are these percentages without some benchmark against which to compare them? Suppose one found that 35 percent of an institution's students reported reading during the previous two weeks a book that was not assigned by their professors. Is that percentage good, bad, or indifferent? Interpretation of this finding would be much more meaningful if one could also inform the institution's faculty and administrators that this percentage was one of the highest (or lowest) rates reported by students at similar colleges and universities. Likewise, suppose colleagues at a college were told that the percentage of students marking some choice or the other had increased 10 points over the previous year. Again, is that finding good, bad, or indifferent? To give this specific finding meaning, one needs a comparison-based benchmark in which to frame it. Furthermore, why assert that a change is relevant unless one can display its connection with learning? In this sense, the comparison to other colleges and universities, or to previous classes at the same institution, provides the benchmark, or frame of reference, necessary for interpreting results.

The principal argument of the "improvement" movement is that it is only the assessment instruments assisting faculty in the classroom that are useful and that faculty are best served if they design and use their own assessments, such as portfolios.

The principal argument of the "improvement" movement is that it is only the assessment instruments assisting faculty in the classroom that are useful and that faculty are best served if they design and use their own assessments, such as portfolios (a recent favorite). According to this argument, standardized tests are not viewed as helpful for several reasons including the primary use in standardized assessments of the multiple choice question format. Many faculty members do not find multiple choice tests authentic and, thus, reject them as inadequate in capturing the teaching and learning experience. While the reliability and validity of scoring of standardized tests and the controlled conditions under which standardized tests are given may permit comparisons, many dismiss this point because of the arguments (noted above) that comparisons are unwarranted, unnecessary, or even politically dangerous. Finally, standardized tests are seen as failing the test of directly assisting faculty in the classroom—leading to our paradox.

Formative assessment alone is not a reliable basis for the improvements that adherents of formative assessment aim to achieve. Without the use of appropriate standardized tests, the assumption that the best way forward is formative assessment focused on single institutions is fatally flawed—because this approach provides no empirical basis on which to make comparisons of promising formative assessments. One-off tests within single institutions alone cannot be reliably interpreted because there is no reliable way to compare the results across time within the institution, across programs within the institution, or across institutions.

To argue that between-institution comparisons and, hence, standardized tests have an important role in assessment is not to say that formative assessments focused on assisting faculty to improve teaching and learning are not needed or are inappropriate. Quite the opposite is the case. Standardized tests should be combined with formative, within-institution assessments. Both are necessary if we are to achieve something like the continuous system of improvement of teaching and learning that many in the academy desire.[6]

[6] One step might be to include more representatives of the measurement-science community on the boards of groups focused on assessment for improvement. Members of that community who would likely impart useful knowledge as well as benefit from such interaction are measurement scientists such as Henry Braun, Boston College; Edward Haertel, Stanford University; Larry Hedges, Northwestern University; Paul Holland, ETS chief research scientist emeritus; Dan Koretz, Harvard University; Robert Linn and Derek Briggs, University of Colorado; and Paul Sackett, University of Minnesota, among others.

3. What is really important is what goes on in the classroom between the teacher and the student, and standardized tests don't capture this.

Yes, this is true. However, there is a growing consensus about the need for reform in undergraduate education that can be characterized as shifting along three dimensions toward (1) a student-centered approach; (2) a case or problem approach in courses and curriculum; and (3) more open-ended assessment instruments.

To achieve such reform, to help them shift their pedagogy, course design, text selection, and assessments, faculty need tools that tell them whether and how much they are improving. We have always needed stronger theories on teaching and learning, but we have yet to develop such theories. In these circumstances, the ability to compare the outcomes of an institution's courses, programs, and overall contribution to student learning outcomes is essential because it gives instructors and administrators a larger arena to study and adapt best practices that will help them improve. Finally, what you assess in large part determines what you teach. It is vital to move beyond multiple choice tests to open-ended essay tests. For example, multiple choice tests may present examples of correlations and causation and then ask students to identify which is correct or ask them to choose whether each is correctly or incorrectly applied. However, responding to such choices passively is very different from asking students to actively critique a case study or to present an argument about data in which correlation and/or causation are misused. In the latter approach, the student must not only recognize the mistake but must also understand where and how the concepts are confused and must explain why the argument fails.

Reform in undegraduate education is shifting along three dimensions:

1. ***A student-centered approach;***
2. ***A case or problem approach in courses and curriculum; and***
3. ***More open-ended assessment instruments.***

4. One-size-fits-all measures to compare institutions are inappropriate.

No single measure can capture completely the complexity of a particular college or university. Interpretations of mean scores on standardized tests at the institution level should be set in the context of multiple indicators collected and analyzed by faculty at the institution itself. The paragraphs below present a four-part framework for how an institution might respond to the initial institution-level scores on a standardized assessment, with examples illustrating how administrators and faculty can benefit from using standardized test scores along with other measures related to student learning. In a basic sense, it does not matter where the institution falls in relation to other comparable institutions the first time it tests. The comparison gives faculty and administrators a benchmark, a signal about where their institution stands. The question, then, is what the faculty and administrators of institutions should do to improve their value added. This leads to the following steps:

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

b. *Analyze results in depth*. While the institution's score signals the place of the college compared to all other colleges administering the higher order skills tests, college administrators and faculty members will want to know more about the relative contributions to that score by colleges (if the institution is a university) or by certain departments or programs (if the institution is a college). For example, which departments or programs are particularly strong or weak contributors to their higher order skills test results?

c. *Audit existing assessments*. There is a saying that faculty value what they measure. Therefore, choosing assessment instruments that accurately reflect what faculty value is critical in determining student and institution performance.[7] Recently, a number of new tests hold the promise of having direct classroom use for improving student performance as well as for providing summative test results. Examples include the tests developed for Carnegie Mellon's Open Education Initiative (OEI), simulations and interactive games,[8] and performance assessments now being developed in large numbers through the partnerships of testing organizations (such as CAE, CBT McGraw-Hill, the College Board, the Educational Testing Service, and Pearson)[9] contracted to develop the next generation of 21st century assessments by two consortia funded by the U.S. Department of Education.[10] These tests all ask students to apply their knowledge to new situations. Education technology promises to enable a much wider variety of tests for future use in the classroom, as well as more sophisticated analyses that use a greater variety of test instruments.[11]

d. *Examine best practices demonstrated to produce good test results* (Sotherland, Dueweke, Cunningham, & Grossman, 2007). Ideally, the most important step is this: Get tests valued by the faculty into the hands of faculty so they can

- use them in their own classrooms, where they have greater knowledge of the strengths and weaknesses of their students;
- develop assessments that challenge students to apply what they know to new situations;
- choose case studies and problems for instructional materials consistent with documents in the assessments rather than drawn from existing content-dominated textbooks;
- adopt a student-centered approach to teaching that calls for more analytic-based writing and diagnostic feedback about how students can improve their performance.

These four steps comprise an early version of what I hope will become a reinforcing system of continuous improvement of teaching and learning. The institution's global score provides a critical signal of the institution's comparative standing, which, in turn, can trigger an internal focus on what correlates with the score. The institution's initial score does not really matter. What does matter is understanding what brought about this result and determining improvement goals for the future.

The institution's global score [representing student learning] provides a critical signal of the institution's comparative standing, which, in turn, can trigger an internal focus on what correlates with the score.

[7] One additional national survey instrument must be cited: the widely used National Survey of Student Engagement (NSSE). Although NSSE is not a direct assessment of student learning, the NSSE initiative pioneered the use of survey data for understanding the best practices associated with improving student learning and has produced an important set of recommendations (Kuh, 2008). See NSSE's annual reports at http://nsse.iub.edu/. See also the Wabash National Study of Liberal Arts Education, which integrates the results of standardized tests, NSSE, and its own surveys to provide in-depth studies of best practices associated with increasing student learning. Reports from the Wabash study may be accessed at www.liberalarts.wabash.edu.

[8] See *Simulation & Gaming: International Journal of Theory, Practice, and Research* for articles that provide examples.

[9] I am Director of the Council for Aid to Education (CAE). For further information about the next generation of assessments see the news announcements at www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse and www.smarterbalanced.org/news/smarter-balanced-awards-pilot-item-and-performance-task-development-contract-to-ctbmcgraw-hill/

[10] For links to reports from colleges and universities applying performance assessments for a variety of instructional purposes, see Benjamin et al. (2012); cf Paris (2011), Roksa (2012), and Sotherland (2007).

[11] The CAAP, the CLA, and the Proficiency Profile have featured a protocol focused on the added value an institution provides to students by comparing freshman and senior test results, controlling for the entering competencies of freshmen. The growth in the number and variety of tests through the next generation of 21st century national programs makes it possible to create new protocols to produce reliable and valid results for large numbers of individual students. This means that it is now possible to use direct measures of student learning more widely as part of student learning outcomes measurement systems and that, therefore, such systems will likely be developed to provide necessary components of attempts to measure the productivity of instruction. See Sullivan, Mackie, Massy, and Sinha (2012) for a recent review of the current problems associated with the measurement of productivity in higher education.

5. Content is what is important in undergraduate education.

Of course content is important, but in today's Knowledge Economy the application of knowledge to new situations is equally—if not more—important. Before the onset of the Knowledge Economy, there was a sense that there was an attainable stock of knowledge and that the job of lecturers was to pour this knowledge into students, who were passive receptacles to be filled to the brim. But we now live in an age where one can "Google" to access facts. It is more important to be able to access, structure, and use information than merely to accrue facts. Recent theories of learning, reflecting the change in emphasis from a focus on content to a focus on higher order skills, are redefining the concept of knowledge. Herbert Simon (1996, p. 4) argues that the meaning of "knowing" has changed from being able to recall information to being able to find and use information. Bransford, Brown, and Cocking note that the "… sheer magnitude of human knowledge renders its coverage by education an impossibility; rather, the goal is conceived as helping students develop the intellectual tools and learning strategies needed to acquire the knowledge to think productively" (2000, p. 6).

Of course content is important, but in today's Knowledge Economy the application of knowledge to new situations is equally - if not more - important.

This does not mean that the content taught in academic majors, for example, is unimportant; they are the fundamental building blocks for both the transmission of knowledge across generations and the creation of new knowledge. But employers also want employees who can navigate the increasing flood of information and come to reasoned judgments about appropriate courses of action.

6. If colleges and universities engage in standardized assessment, the results will be used by state and federal authorities to control and punish institutions that score low on arbitrary and capricious indicators.

Consider the response of public authorities to the production of the other main public good that universities produce: research. Although there is great variation in the amount and quality of research produced across universities, there is consensus about the considerable resources in infrastructure and academic talent that universities need in order to produce quality research. These understandings, crystallized in the Bush Report (1945), have developed over the past several decades. Once similar factors of production are known, it is likely that political leaders will embrace solutions to improve student learning by responsible higher education leaders as well. Initial comparisons of student learning outcomes at the institution level are important signals to faculty and administrators about how well they are doing. More important is what institutions actually do to improve. Higher education institutions are much more complex and autonomous than elementary and secondary schools. Because of this, higher education leaders can pave the way in deciding what is appropriate to report publicly about student learning. That is what two national associations of higher education, the Association of Public Land-grant Universities (APLU) and the American Association of State Colleges and Universities (AASCU), are attempting to do now.[12] Developing assessments that are embraced by faculties is a necessary step in any call by external forces demanding accountability. Only when faculty members accept assessment instruments as effective aids to their classroom success will the higher education community, and the groups that hold the higher education community accountable, remove the blinders and move past these false arguments.

[12] See http://www.voluntarysystem.org/index.cfm

7. It is impossible to conceptualize a framework for standardized assessment that faculty and external authorities would agree to publicly report.

As in the case of the peer review of research, comparative assessments of learning should be designed to be objective characterizations of institution-level performance on student learning outcomes. As such, they should provide evidence that is judged reliable and valid, and should also have strong face validity.[13] Such assessments that meet the measurement-science requirement of minimum standards of reliability and validity offer a powerful reality check for institution-based formative assessment. But the organization that does the testing should not be the one to make the results public. The testing organization should report assessment results for the institutions it tests to those institutions only. Otherwise, why would an institution, department, or program permit comparative-based testing, which, we argue, plays a critical role in making formative assessment more systematic?

Unless, or until, standardized assessments judged by faculty as authentic are widely accepted and widely supported, effective design alternatives will not be developed, and the status quo will prevail.

Conclusion

Because it is possible to learn from other institutions that are demonstrably doing well in teaching and learning, there is no intellectual argument not to do so. The stakes are too high. Peer review anchors the system of continuous improvement that advances scholarship and scientific research, and we should adopt its principles with respect to student learning outcomes. We may never achieve in the realm of teaching and learning the clearly positive results that we have achieved in scholarship and creative activity from peer review. Surely, however, we can use the basic strategy suggested here to move the subject of teaching and learning to a much more evidence-based approach in which verifiable best practices are continually adjusted or changed and improvement of student learning demonstrably occurs. The main reason for the relatively little progress that we have achieved in assessment in higher education is institutional inertia. All organizations, including universities and colleges, have set up protocols and decision rules to undertake certain services deemed important for private or public reasons. Institutions, like the individuals that inhabit them, tend to continue their familiar behavior patterns and to resist developing new practices because change requires decisions, and decisions involve risk. Unless, or until, standardized assessments judged by faculty as authentic are widely accepted and widely supported, effective design alternatives will not be developed, and the status quo will prevail. We are not there yet, but the essential components now exist to realize such a design when exogenous and endogenous forces come together to render it real.

[13] If an instrument or method has "face validity," it generates results that make sense to lay constituents. It "…pertains to whether a test looks valid to the examininees" (Anastasi, 1988).

References

Anastasi, A. (1988). *Psychological testing*. New York, NY: Macmillan.

Benjamin, R. (2008, November/December). The case for comparative institutional assessment of higher order skills. *Change: The Magazine of Higher Learning*, *40*(6), 50–55.

Benjamin, R. (2012). *The new limits of education policy: Avoiding a tragedy of the commons*. London, UK: Edward Elgar.

Benjamin, R., Klein, S., Steedle, J., Zahner, D., Elliot, S., & Patterson, J. (2012, May). *The case for generic skills and performance assessment in the United States and international settings* (CAE Occasional Paper No. 1). New York, NY: Council for Aid to Education.

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

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (Expanded edition). Washington, DC: The National Academies Press.

Bush, V. (1945, July). *Science: The endless frontier*. Washington DC: United States Government Printing Office.

Klein, S., Lui, O. L., Sconing, J., et al. (2009, September 29). *Test validity study (TVS) report*. Washington DC: Fund for the Improvement of Postsecondary Education. Available at http://www.voluntarysystem.org/docs/reports/TVSReport_Final.pdf

Kuh, G. D. (2008). *High-impact educational practices: What they are, who has access to them, and why they matter*. Washington, DC: Association of American Colleges and Universities.

Paris, D. C. (2011). *Catalyst for change: The CIC/CLA Consortium*. Washington, DC: The Council of Independent Colleges.

Roksa, J. (2012). *An analysis of learning outcomes of underrepresented students at urban institutions*. Washington, DC: The Council of Independent Colleges.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.

Smarter Balanced Assessment Consortium. (2012, April 16). *Smarter Balanced awards pilot item and performance task development contract to CTB/McGraw Hill*. Retrieved from www.smarterbalanced.org/news/

Sotherland, P., Dueweke, A., Cunningham, K., & Grossman, B. (2007, Spring). Multiple drafts of a college's narrative. *Peer Review*, *9*(2), 20–23.

Steedle, J., Kugelmass, H., & Nemeth, A. (2010, July/August). What do they measure? Comparing three learning outcomes assessments. *Change: The Magazine of Higher Learning*, 42(4), 33–37.

Stevens, A. (2010). *Summary of mission statement research*. Unpublished manuscript. New York, NY: Council for Aid to Education.

Sullivan, T. A., Mackie, C., Massy, W. F., & Sinha, E. (Eds.). (2012). *Improving measurement of productivity in higher education*. Washington, DC: The National Academies Press.

U.S. Department of Education. (2010, September 2). *U.S. Secretary of Education Duncan announces winners of competition to improve student assessments*. Retrieved from www.ed.gov/news/press-releases/

Wabash National Study of Liberal Arts Education. Retrieved from www.liberalarts.wabash.edu

Wagner, T. (2008). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need—and what we can do about it.* New York, NY: Basic Books.

Demonstrating and Improving Student Learning: The Role of Standardized Tests

Margaret A. Miller

In 1986, the Virginia legislature issued Senate Document #14, which mandated assessment for all the public colleges and universities in Virginia. A study resolution introduced the prior year had concluded that for assessment to serve its chief purpose of improvement, each campus's faculty would design their assessments to conform to the learning expectations of their particular programs and that unique institution. Accountability would reside in the assurance that every campus was keeping track of its results and making improvements when they were called for.

This assumption held for about a decade, during which I read (as chief academic officer of the coordinating board) every assessment report written in the state. Even so, at the end of that time, I couldn't answer a deceptively simple question a legislative staffer asked me: "So, how are we doing?"

Accountability had another face, it seemed—public representatives wanted to know not only that we were keeping track of our results but what those results were. Since the assessment program couldn't supply an answer to the latter question, legislators turned to measures of performance outcomes. But learning, the chief goal of higher education, couldn't be one of those outcomes because we had no standardized and reliable measures of it.

It turns out that the question "How are we doing?" can only be answered with reference to a prior question: "Compared to what?" Is an 80% pass rate on a home-grown exam good news or bad? We can't know until we can compare it to the results of another, similar group of students who take the same test.

Roger Benjamin's paper makes another, even more radical, point: that even for purposes of improvement, standardized instruments are important. And the reason is the same as it is on the accountability side: Without the capacity to compare one institution to another, it is impossible to make meaning of the results. The "how are we doing" question is as central to campus-based assessment for improvement as it is for assessment done for the sake of accountability.

Atul Gawande, in an article in *The New Yorker* called "The Bell Curve" (December 6, 2004), made the same point in writing about the success of cystic fibrosis treatment centers. People are apt to judge the quality of medical centers the same way they do colleges: by the prestige and reputation they have acquired. But these don't have any necessary connection to the results they get. There is a bell curve among institutions: Some, treating the same disease or educating similar students, get better results than others.

This is distressing news to doctors and professors alike. As Gawande says, "The bell curve . . . contradicts the belief nearly all of us have that we are doing our job as well as it can be done." And there are no quarterly reports or win-loss records to serve as reality checks, he points out. Standardized testing is our reality check, the antidote to groundless self-congratulation.

The question "How are we doing?" can only be answered with reference to a prior question: "Compared to what?"

The Red Herrings

Benjamin takes on seven red herrings about assessment in the paper, countering them with effective arguments. Some of his responses resonate especially strongly with my experience.

Standardized vs. Home-Grown Test Quality

Benjamin emphasizes the quality of the newest generation of standardized measures: Not only are they designed by testing-and-measurement specialists, carefully field tested, and checked for validity and reliability, but the latest ones replace multiple choice questions with authentic performance tasks that have a considerable amount of face validity. Computer scoring has finally made this form of testing feasible—although arguably, faculty learn a great deal in scoring their own students' work on standardized tests, even if the task is time consuming.

Judging from the assessment reports I read over the years, very few professors have the knowledge and skills to design assessments of student work that are reliable and valid. This is not to fault the faculty: Very few of us were given the relevant training. But even those who do possess that knowledge rarely spend the time and energy required to make their assessments as good as they can be. This may be because faculty are rarely rewarded for taking time from their busy schedules to do such work. Sometimes too it is a consequence of assessment's not being taken seriously. I remember one program in Virginia that reported that 83% of its students had passed its assessment exam with the grade of "sundae with a cherry on top." Rarely is the resistance to assessment that blatant, but it nevertheless persists.

The Importance of Multiple Measures

Benjamin argues that standardized, summative assessments should be paired with localized formative assessments. I agree. Standardized assessments can raise flags, but they rarely tell faculty exactly where in the curriculum the problems lie or what to do about them. Moreover, local measures target the issues that matter to the faculty and use methodologies with which they are familiar. Finally, embedded assessments can be used in the teaching and learning process—feedback is, after all, essential to learning. A side benefit of embedded assessment, moreover, is that it avoids the problems of student motivation and testing aversion.

Having given my caveats about quality above, however, I believe that disciplinary communities should begin to discuss and share best practices in assessment. Recognizing this problem, the Council for Aid to Education has developed a program called CLA in the Classroom, which instructs faculty how to create performance tasks for classroom use. Scholars of teaching and learning are also experimenting with assessment strategies that will increase our understanding of how students learn. This work needs to be encouraged and rewarded.

Another reason to have multiple measures of learning is because learning assessment, to the despair of measurement scientists, is unavoidably messy. It deals with people, we are rarely able to set up double-blind tests, and meeting the gold standard of research is almost never possible. So we need to look for arrows that all point in the same direction.

Embedded assessments can be used in the teaching and learning process—feedback is, after all, essential to learning.

The Value of Applied Knowledge and General Intellectual Skills

Insofar as scholars of my generation thought about the purposes of teaching at all, we tended to assume that it was primarily about the transmission of knowledge. This was brought forcibly home to me on a visit to Hungary. Lacking a full complement of textbooks and primary sources, professors there spent many classroom hours reading books and articles to their students, much as medieval scholars imparted the contents of precious and rare manuscripts to their students.

Undoubtedly, mastery of a common set of basic concepts and crystalized knowledge is necessary element of disciplinary mastery. A physics student who doesn't have down cold the force concept inventory (the most basic concepts in Newtonian physics) or a biology major who is ignorant of evolutionary theory would not have the foundations on which to build further understanding of their fields. But in the U.S. we have textbooks and primary materials galore, and anyone with a smartphone can look up facts on the run. Also, some content is quickly rendered out of date. Even a couple of decades ago, both the then-Big-Six accounting firms and a group of manufacturers in Virginia were saying that trying to equip students with permanent and complete content knowledge was a fool's errand for academia.

Our job, instead, is to teach students how to "find and use information," as Benjamin says, and how to update it continuously. This points to the importance of general intellectual skills. Too few of us think about how our disciplines discipline the minds of students, which means that not only in the majors but in general education we continue to think that our job is to impart knowledge. We thereby miss the point. The abilities to think critically, communicate effectively, and solve problems are so key to the kind of intellectual nimbleness that will be required of our graduates that faculty members who don't include them among the learning goals in their classrooms are arguably guilty of educational malpractice. And those skills should have pride of place in our assessment of learning.

The abilities to think critically, communicate effectively, and solve problems are so key to the kind of intellectual nimbleness that will be required of our graduates that faculty members who don't include them among the learning goals in their classrooms are arguably guilty of educational malpractice.

The Importance of Using Results

Trudy Banta and Charles Blaich, in an article in the January/February 2011 issue of *Change* magazine, report on their depressing failure to find institutions that are "closing the assessment loop" by using the results of assessment to make improvements in their programs. "This is difficult enough with locally developed measures," they say. But "adding the need to interpret nationally standardized test scores and connect them with local programs and teaching approaches exacerbates the difficulty of the task" (p. 22).

Benjamin describes a four-step process for doing assessment that culminates in "examining best practices demonstrated to produce good test results" so that faculty can imitate those successful practices. This is not only a lot of work; it also runs up against the core faculty values of originality (they shouldn't imitate their colleagues) and academic freedom (their classrooms are inviolate). We need to reconsider what we mean by these terms and what limits we should put around them.

Caveat

I do take exception to one of Benjamin's conclusions: that testing organizations should not make the results of standardized assessments public. They could make such an agreement with institutions a condition of participation, as the Community College Survey of Student Engagement has done. Very few people or institutions welcome this kind of exposure, of course, in either health or education. Beyond the difficulties of figuring out what to measure, what instruments and metrics to use, and how to find the time and resources to do the work, there is the possibility of finding one's results to be mediocre or worse.

Don Berwick's argument for openness, as paraphrased by Gawande, is that it is likely to "drive improvement, if simply through embarrassment." Indeed, we have seen its salutary effects on quality in assessing student work. Electronic portfolios that are available to family and friends and that may even be used for job hunting or culminating projects presented to colleagues and industry representatives are likely to elicit from students their very best efforts.

More important, however, we need to make results public so that we can learn from each other. Some institutions are doing their work better than others, and the less successful institutions should be able to take a lesson from them. A point that Benjamin makes repeatedly is that "because it is possible to learn from other institutions that are demonstrably doing well in teaching and learning, there is no intellectual argument not to do so." But this is possible only when we know who is doing well and who is not. Making results public provides those who are not doing well both with a motivation to do better and with the knowledge of whom to turn to in order to find out how the most successful get their results.[1] Meanwhile, the most successful are encouraged to keep up the good work.

The problem/good news is that if higher education is anything like health care, the target will move. Once the results on the cystic fibrosis centers were made public, the best ones got better faster than the less successful ones did, so that mediocre centers that had improved continued to reside in the middle of the pack. It's much like the Olympic Games. Times get faster every year, so that silver medal winners may increase their personal best from one Olympics to the next, only to find themselves with the bronze instead of the gold medal they had aspired to. Still, the moderately good have motivation to improve. While they might settle for barely making it to the platform, no one wants to be the last one across the finish line.

Acknowledgement: I would like to thank David Shulenburger and Christine Keller for their helpful information about and comments on the VSA.

References

Banta, T., & Blaich, C. (2011). Closing the assessment loop. *Change: The Magazine of Higher Learning, 43*(1), 22-27.

Gawande, A. (2004, December 6). The bell curve. *The New Yorker.* Retrieved from http://www.newyorker.com/archive/2004/12/06/041206fa_fact

[1] Actually, the chief lesson for the cystic fibrosis treatment and teaching effectiveness may turn out to be the same: Think hard about your patients/students, push them, improvise to get better results, insist that the most successful improvisations be uniformly adopted, and keep exercising the capacity "to learn and adapt."

The Voluntary System of Accountability (VSA) has committed to presenting comparable learning results based on standardized testing (For NILOA's assessment of the VSA, go to http://www.learningoutcomeassessment.org/documents/VSA.htm). NILOA's evaluation found that only about half of the 320 participating universities have posted their results (although those that have not will be asked to publicly report on the reasons why they were unable to do so). Moreover, most of the institutions that have administered the tests have not found the results to be useful in improving curricula—although whether this is the cause or effect of the failure to close the assessment loop is a key question.

Nevertheless, the participating institutions have decided that the VSA should be continued—but with expanded reporting options: The AAC&U VALUE rubrics and the GRE General Test are now on the list of approved instruments. The VSA board will also allow the reporting of raw scores (within a normative or comparative distribution), given the instability of value-added scores over time; at the same time, more contextual information will be required. To see the VSA's Preliminary Outline for Expansion of Student Learning Outcomes Reporting, go to http://www.voluntarysystem.org/docs/reports/VSABoardDecisions_060712_final.pdf.

Getting Serious About Assessing Authentic Student Learning

Terrel L. Rhodes

There are many adages, often said with a chuckle, that change in the academy is slow and painstaking. On the whole, this is true. Academics and their institutions resist the fad, the easy answer, the quick fix that seems to pervade much of modern society and the media-driven public discourse around improving student performance in our colleges and universities. Faculty are socialized to challenge, debate, test, generate, and evaluate evidence to support claims. Questions about student learning—what it should be and how to assess it—are no different.

Faculty are socialized to challenge, debate, test, generate, and evaluate evidence to support claims. Questions about student learning—what it should be and how to assess it—are no different.

Roger Benjamin's "Seven Red Herrings" brief favors standardized tests. I came away from reading it with a hollow feeling. The red herrings, taken together, create a straw person, which is then rejected. In the rest of this paper, I argue that assessment of student learning in practice has moved beyond the traditional modes and arguments for assessment. I illustrate this in the faculty-led cultures of assessment emerging on campuses focused on student work and faculty guidance and in the multiple modes of representing demonstrated levels of the quality of student learning using new assessment tools that faculty across the country have already developed and continue to develop.

Limitations of Using Standardized Tests to Assess Student Learning

I fully subscribe to and endorse AAC&U's position not to accept the premise that standardized tests are sufficient to what students know and can do as a result of their college experience. "AAC&U does believe, however, that standardized tests can supplement curriculum-embedded assessments when they are used with appropriate professional standards and cautions."[1]

It is the case that some faculty resist standardized tests; but many others embrace them. Even so, the more important point is that we don't need a new and improved standardized test or suite of tests. We need something to evaluate student learning that actually reflects the demands of the 21st century rather than the 20th or even the 19th century. For too long we have relied upon a family of standardized tests for college admissions and their progeny for college achievement, knowing full well their many limitations, including being correlated so strongly with family income that they contribute to sustaining our nation's serious and ongoing inequities.

Standardized tests are used extensively in K–12 education. One of the lessons we have learned with stark clarity from K–12's experience is that when decisions are made based on standardized test results of a very few learning outcomes, as we have done in the schools in this country, virtually every other critical learning outcome disappears from practice. The evidence from faculty and employers alike is unanimous. Our graduates need more than the limited range of competencies easily measured by standardized tests; they must have a broad array of essential learning outcomes if they are to be successful and vibrant contributors to the civic fabric of our country, to the global community, and to an interdependent economy.

[1] AAC&U, 2007.

As my colleague, Carol Geary Schneider, president of the Association of American Colleges and Universities, stated in early 2009 in *The Proof Is in the Portfolio*, "We are scholars and we are educators. As scholars, we need to mobilize the already abundant evidence showing why *narrowly focused* standardized tests are misaligned with the way knowledge is actually put to work in the twenty-first century context. As educators, we need to move beyond the reactive mode provoked by the Spellings barrage and help society get ahead of the curve on forms of assessment that can actually drive higher achievement."[2]

As scholars, we need to mobilize the already abundant evidence showing why* narrowly focused *standardized tests are misaligned with the way knowledge is actually put to work in the twenty-first century context.

We now have resources and modes for assessing student learning that reflect our current century, namely e-portfolios and rubrics. Whether institutions have taken the leap and are using e-portfolios or not, most are using rubrics for assessment and/or grading student achievement. In a recent survey conducted by the Association for Authentic, Experiential, and Evidence-Based Learning (AAEEBL), over half of U.S. higher education institutions are using e-portfolios.[3] Rather than relying solely on one-off tests with few consequences for students and no connection to the curriculum, assessment of work in portfolios focuses on what students produce as a result of faculty and staff assignments embedded in the curriculum and co-curriculum. Would any employer retain or promote an employee based on their performance on a one-time test and ignore the actual quality of the work they deliver day to day on their job? Why, then, would colleges and universities, or policy makers settle for low-stakes tests rather than a fuller set of high-stakes, authentic work judged to be important by the faculty and staff who are responsible for the educational outcomes of a degree or certificate?

Using VALUE Rubrics to Assess Student Learning

Benjamin rightly points to the need for external validation and reliability. The development of common rubrics allows us to address these issues. Supported by a grant from the Fund for the Improvement of Post-Secondary Education, the Valid Assessment of Learning in Undergraduate Education (VALUE) project engaged over 100 faculty from across the country, representing all types of two- and four-year institutions and disciplines, in a proof-of-concept exercise demonstrating that faculty broadly agreed on what student learning looked like when they saw it. They also agreed that the key elements of the LEAP Essential Learning Outcomes[4] could be and needed to be articulated at progressively more sophisticated and accomplished levels. By examining rubrics developed by faculty at many different institutions, reports from centers and institutes focused on researching outcomes such as critical thinking, writing, and creativity, and from experts in the field, these teams of faculty drafted rubrics for 15 areas of learning for college graduates deemed essential by faculty, policy makers, and employers.

Once drafts of the VALUE rubrics were completed, over 100 two- and four-year institutions tested the rubrics with their own students' work and provided feedback from the respective faculty and staff so that the rubrics could be revised for usability and clarity. Each rubric passed through two to four rounds of testing and revision before the VALUE rubrics were released in fall 2009, to be used by faculty and staff, free of charge. A reliability study was conducted before the conclusion of the project involving 40 people drawn from faculty and staff who had not been involved with the project, school teachers, employers, and nonprofit administrators. For two days, this group engaged in a calibration effort on the use of rubrics. They examined three of the VALUE rubrics and used them to assess a set of student e-portfolios from a broad range of institutions including LaGuardia Community

[2] Schneider, 2009.

[3] Brown, Chen & Jacobson, 2012.

[4] AAC&U, 2005

College, Spelman College, and the University of Michigan. The results demonstrated that a diverse set of individuals who knew nothing about rubrics and e-portfolios or one another could reach methodologically acceptable levels of agreement on the quality of student learning exhibited in the work produced through their respective curricula using the VALUE rubrics.

An additional reliability study, in 2011, focused on whether disciplinary preparation of faculty would make a difference in the use of rubrics. Forty faculty members from four traditional disciplinary divisions—social sciences, humanities, science, and the professions— again utilized three VALUE rubrics and a set of student work samples to assess quality of student learning. The results revealed minimal standard deviation differences in the assessment of student quality of performance based on the disciplinary background of the reviewers or the students.[5]

Since the end of summer 2010, when tracking data began to be collected, more than 3,000 different institutions and more than 11,000 individuals downloaded and have been using one or more of the VALUE rubrics. Institutions that are using the VALUE rubrics have reported that, once calibration sessions are held with their faculty, they achieve the industry-accepted standard of .8 or higher agreement on the quality level of student performance.[6] Several consortia of institutions are using VALUE rubrics and student work across institutions to calibrate their expectations for quality learning as students transfer from one institution to another. Faculty and institutions clearly are viewing the VALUE rubrics as valid measures of student performance for assessment purposes.

The major e-portfolio commercial and open-source providers have all incorporated the VALUE rubrics into their e-portfolio platforms. E-portfolio technology is being merged with institutional Learning Management Systems to marry the collection of student work from across the curriculum and co-curriculum into a seamless framework for combining faculty assessment judgments, student reflection and self-assessments, and institutional data on students into a single repository that can be used for informational and accountability reporting.

Assessment Should Lead to Improved Student Learning

Some faculty members do resist assessment, as Benjamin points out, because they view assessment as something that the institution does primarily to respond to demands for accountability. But such faculty do not represent the majority. As Benjamin also notes, we should be investing in assessments that yield data that faculty can use to improve student learning. Connecting assessment to the day-to-day work of the faculty and staff through e-portfolios and rubrics accomplishes this goal. This formative feedback also allows students to have a clearer idea of their own learning strengths and weaknesses in the context of the dimensions or criteria of learning reflected in the rubrics used by faculty.

The results of using rubrics and portfolios of student work also can be aggregated for programmatic and institutional reporting through sampling students and their work or examining whole populations. Indeed, we create another red herring by separating assessment for improvement (formative assessment) from assessment for accountability (summative assessment). Institutions across the country are using their resources efficiently and effectively through portfolio and rubric assessment to accomplish both of these desired results in one process.

Formative feedback also allows students to have a clearer idea of their own learning strengths and weaknesses in the context of the dimensions or criteria of learning reflected in the rubrics used by faculty.

[5] Finley, 2011.

[6] Finley, 2011.

Another clarion call in recent years has been for greater transparency of information in higher education around student learning, not for a specific type of information, i.e., standardized tests. Well, we have the ability to aggregate general learning achievement through e-portfolios benchmarked against broadly shared expectations or standards for learning, such as rubrics that articulate key characteristics of quality performance and that provide a set of examples of the genuine work students are producing, allowing anyone interested to judge the quality of learning occurring in colleges and universities. Given this prospect, do we really need tests that allow us to nationally compare institutions on the basis of standardized tests divorced from the actual required curriculum?

Approximately 80% of college-goers are place-bound[7]—that is, having little choice of where to attend college, they go to the nearest affordable institution. This means that most students will benefit from knowing what quality learning expectations look like at their prospective institution, what students actually produce who attend the institution, and how many perform at that level. A growing number of higher education institutions are engaging their advisory boards of alumni and employers with high success in using rubrics with student portfolios or student work samples to gauge better how well the curriculum and faculty are preparing students for postgraduation. What better transparency than a digital ability to state the desired quality of learning (a rubric) and an ability to see a student's actual academic work performance to determine achievement (a portfolio)?

Faculty embrace assessment when they have had a direct hand in developing the assessment, when the information provided is actionable immediately, and when the assessments are aligned with the curriculum and assignments that comprise the student's learning experience.

Conclusion

In short, we have to move beyond the debates about whether standardized tests are reliable, have acceptable psychometric properties, and can be proxies for a limited number of learning outcomes. Even with new, improved performance-based tests, we still have less than what we need to make good judgments about student attainment. The undergraduate experience is replete with authentic performances. Why do we need measures that are isolated and divorced from the curriculum and faculty? Why do we want to keep settling for a very limited, few learning outcomes, when all of the evidence we receive from faculty, employers, and the media is that our students must have competencies in a much broader set of essential learning outcomes, such as civic learning, teamwork, intercultural knowledge and understanding, ethical behavior and so forth? Why do we limit ourselves to assessing student learning through text exercises when jobs and life require application of learning in real-world contexts, working with diverse others rather than solely as an individual, employing multiple media, researching and communicating in multinational, multicultural teams distributed around the globe in real time?

Tools such as e-portfolios capture the impact of the curriculum and co-curriculum in its many facets, modes of learning and media; the students' best work in the form of graded assignments, which are the motivational coin of the realm for students as well as the articulated expectations of quality performance shared by faculty across the country. Yes, more research is needed. But in the meantime, institutions are beginning to examine the costs of standardized tests compared to using e-portfolios and rubrics, and they are finding the latter to be less expensive. A few institutions are using e-portfolios and quality of demonstrated performance as graduation requirements— truly high-stakes performance metrics. Many campuses know that using rubrics and student work helps generate conversations across the silos of higher education, moving student learning from isolated, individual work to work that is shared across courses, faculty, divisions, and even institutions.

Rubrics and e-portfolios are providing students with formative and summative feedback so students can develop the ability to reflect upon and judge their own quality of learning, something standardized tests do not permit them to do at the same level. It is hard to imagine a better outcome for our graduates for life beyond college.

[7] WICHE, 2005.

Faculty assess student performance every day. Faculty embrace assessment when they have had a direct hand in developing the assessment, when the information provided is actionable immediately, and when the assessments are aligned with the curriculum and assignments that comprise the student's learning experience. It's time to move beyond the timeworn arguments to new assessment approaches that are sensitive to and respond to the needs of the world in which we live and the lives that our students, faculty, and institutions are creating.

References

Association of American Colleges and Universities. (2005). *Liberal education outcomes: A preliminary report on student achievement in college*. Washington, DC: Author.

Association of American Colleges and Universities. (2007). *College learning for the new global century.* Washington, DC: Author.

Brown, G., Chen, H. L., & Jacobson, J. (2012, June). *ePortfolios changing the learning context: The AAEEBL survey report 2011* (The AAEEBL Learner, Vol. 3. No.3). Association for Authentic, Experiential, and Evidence Based Learning. Retrieved from http://www.aaeebl.org

Finley, A. (2011, Fall/2012, Winter). *How reliable are the VALUE rubrics? Peer Review, 13*(4)/*14*(1), 31–33.

Schneider, C. G. (2009). The proof is in the portfolio. *Liberal Education*, *95*(1).

Western Interstate Commission for Higher Education. (2005, May). *Student migration: Relief valve for state enrollment and demographic pressures* (Policy Insights). Boulder, CO: Author.

It's time to move beyond the timeworn arguments to new assessment approaches that are sensitive to and respond to the needs of the world in which we live.

Making the Case Against -- One More Time

Trudy W. Banta and Gary R. Pike

Between 1986 and 1994, Tennessee's performance funding initiative gave us the opportunity to assemble a unique body of work aimed at understanding the properties of standardized tests of generic skills. Our discoveries while working together at the Center for Assessment Research and Development at the University of Tennessee, Knoxville (UTK), convinced us that the state of the art of measurement is not sufficiently advanced to support the use of standardized tests of generic skills for making comparisons among institutions. Roger Benjamin and NILOA provide yet another opportunity for us to share what we have learned about those tests.

For several reasons, these may be considered the worst of times for advocating the use of such tests as the Collegiate Assessment of Academic Proficiency (CAAP), the Collegiate Learning Assessment (CLA), and the Proficiency Profile (PP) to assess generic knowledge and skills. Beginning in the fourth quarter of 2011, events began to unfold that call into question the wisdom of employing these instruments to assess student learning outcomes in U.S. colleges and universities—and especially to compare institutions.

For example, in November 2011, the Council of Independent Colleges (CIC) issued a report based on the experiences of faculty at 47 colleges and universities in administering the CLA to their students and coming together in a series of virtual and face-to-face meetings over the space of three years to discuss the use of the CLA as a catalyst for improving learning on these campuses (Paris, 2011). Many of the participants reported having difficulty interpreting the CLA findings, and the final report concludes, "The CLA results might not be immediately or directly connected to program or pedagogy," and in fact "…there is no clear line between direct and indirect impacts of the CLA on Consortium institutions" (Paris, 2011, p.28).

Then, in March 2012, a comprehensive evaluation of the Voluntary System of Accountability (VSA) by a NILOA team was completed. An important conclusion was stated, "the standardized tests of student learning originally approved for inclusion in the pilot lack credibility and acceptance within a broad sweep of the higher education community which, in turn, serves to undermine institutional participation in the VSA" (Jankowski et al., 2012, p.3). Wasting no time, the staff at APLU and AASCU who oversee the VSA convened a group of methodologists who, at the end of May, recommended using several alternatives to the CAAP, the CLA, and the PP to report student learning outcomes on the VSA website. These alternatives include scores for graduating seniors/alumni on the Graduate Record Exam General Test and student ratings based on AAC&U's VALUE rubrics for Written Communication and Critical Thinking (email from Christine Keller, May 31, 2012).

Against this backdrop, Roger Benjamin offers his defense of standardized tests of generic knowledge and skills. We comment here on three themes that run through several of Benjamin's points:

(1) There are generic outcomes of college that can be assessed;
(2) Standardization in assessment is possible and valuable; and
(3) Making comparisons among institutions will lead to improvement in teaching and learning.

Three themes run through several of Benjamin's points:

1. *There are generic outcomes of college that can be assessed;*
2. *Standardization in assessment is possible and valuable; and*
3. *Making comparisons among institutions will lead to improvement in teaching and learning.*

Generic Outcomes

In a footnote to his introduction, Benjamin defines "good" standardized tests as those that measure college students' *generic knowledge and skills*. That qualifying language is needed throughout the text because there are many excellent standardized tests that measure student achievement in specific areas, including the GRE Advanced Tests in Major Fields and licensing exams in professional fields like pharmacy and veterinary medicine. In fact, skills like written communication, problem solving, and analytic reasoning are learned—and assessed—best as they are **applied in a discipline.** We want and need physicians and teachers who can write and solve problems in their fields and bring the knowledge and perspectives of their disciplines to team problem solving where various disciplines are represented. Testing generic skills developed in college most effectively will use **tests focused on the content of a discipline**.

Testing generic skills developed in college most effectively will use tests focused on the content of a discipline.

Several years ago, the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) developed a discipline-specific measure of critical thinking and problem solving (see Pike, 2001). The test's developers argued that content knowledge was a critical element in the ability to think critically and solve problems (Herl et al., 1999). In a series of studies the test developers evaluated the reliability and validity of the test. An important finding was that prior domain-specific knowledge had a significant (positive) effect on critical thinking and problem solving. These findings have two implications for the assessment of generic critical thinking and problem solving skills.

First and foremost, the CRESST studies raise questions about whether critical thinking and problem solving are *generic skills*. Herl and his colleagues argued that in order to be an effective critical thinker/problem solver, one must have knowledge about the issue at hand (Herl et al., 1999). Likewise, Baird's (1988) review of research on the assessment of generic outcomes concluded that critical thinking and problem solving skills are unlikely to exist free of any context or background. He concluded, "But if we look for evidence about critical thinking and problem solving within discipline or program areas, the results will be much more acceptable and meaningful to faculty. They will have clearer and more specific educational implications and, thus, should lead to appropriate changes of emphasis in courses" (Baird, 1988, p. 53).

A second implication of the results of the CRESST studies is that even if generic critical thinking and problem solving skills exist, developing unbiased measures of these constructs may be impossible. No test item involving scenarios is truly content free. For example, majors in business and engineering will be more comfortable than those in the humanities and the arts in addressing a scenario in which the test taker must decide whether or not to recommend purchase of a particular plane. Pike (1989b, 1990b) examined the ACT College Outcome Measures Program (COMP) Objective Test for evidence of differential item functioning. He found that the test items did not evidence differential item functioning, but the scenarios upon which the items were based did function differently across groups of students. Likewise, Chatman (2007) found that students in different disciplines reported substantially different educational experiences and outcomes. He concluded that generic institutional performance measures may be confounded by disciplinary differences and should be interpreted with caution.

Standardization

Benjamin believes that standardized testing is essential for systematic assessment in higher education. We agree that consistent administration and scoring of assessments are important; however, we question whether such consistency exists today and whether it can exist in the future. For example, there is evidence that standardization, even as related to current administration of the CLA, is not assured. The nature of the sample of students taking the test varies enormously—from those who simply volunteer in response to a mass e-mailing to those who are paid to participate to those who are enrolled in a senior seminar and take the test during a class period. In one state, the CLA has been given only to honors students!

There is evidence that standardization, even as related to current administration of the CLA, is not assured.

Even if samples of test takers could be standardized across colleges and universities, issues of motivation will remain. Motivation to do well on the test also varies enormously, from those who just show up because they are asked to do so to those who are given a vested interest such as earning extra credit in a course or ensuring that their college looks as good as possible in a comparison. In an evaluation of standardized tests at the University of Tennessee, Knoxville, Pike (1988, 1989a, 1990a) found that student motivation was the second-best predictor of students' scores on standardized tests, behind measures of entering ability/aptitude (i.e., ACT and SAT scores). Sundre (2009, vii) noted that motivation to perform well on tests is likely to be particularly low when the "results hold no personal consequences for the students we ask to complete the tasks." In a recent study, Steedle (2010) found that motivation was significantly related to students' CLA scores, but not to institutional mean CLA scores. This finding appears to be due, in large part, to the fact that there was little variance in mean motivation scores in this study. And if students' motivation is related to their test performance, as Steedle and others have found, how can we have confidence in institutional means, which are based on students' scores?

Finally, standardization does not require a national test. Faculty at many institutions have developed their own measures that "are administered and scored in a standardized manner," to use Benjamin's words. For example, following the lead of faculty and staff at Johnson County Community College, faculty at several other community colleges, including Butler County (Speary, 2002) and Western Wyoming (Renz, 2012, in press), community colleges have developed standard procedures for collecting and analyzing student artifacts using faculty-designed rubrics.

Making Comparisons

Our studies and experience convince us that scenario-based test items cannot be content-free. This means that some disciplinary majors will be advantaged and others disadvantaged by the content of tests of generic skills. This, in turn, means that some campuses will be advantaged and some disadvantaged by the mix of majors on the campus and especially by the sample of majors turning up to take the test. What if the scenarios on the test seem most familiar to business and engineering majors, but either no engineering programs are offered on the campus, or even if they are, the engineering students don't bother to report for testing? Will the sample of English and art majors who did report for testing represent the institution as well as would have been the case if the engineering students had taken the test? Jamestown College participated in the CIC study involving 47 institutions. During the first year of testing there, the seniors in nursing and a few other majors were not able to take the CLA due to other commitments. The results were disappointing to the faculty.

The following year steps were taken to ensure that a more representative sample of seniors was tested, and the results were much improved (Paris, 2011, p. 16).

Benjamin says that faculty need to know where the achievements of their students stand in comparison with their peers at other institutions. But given the problems ensuring standardization just mentioned, how can faculty feel confident about the validity of any comparison that might be made on the basis of scores (or value-added statistics) on the standardized tests of generic knowledge and skills? Also, given that institutions may be able to compel students to participate in standardized assessment, but cannot ensure that students will devote the quality of effort to perform their best, institutions may be placed in the perverse position of having high stakes assessments for institutions that are based on low-stakes assessments for participating students.

Given that institutions may be able to compel students to participate in standardized assessment, but cannot ensure that students will devote the quality of effort to perform their best, institutions may be placed in the perverse position of having high stakes assessments for institutions that are based on low-stakes assessments for participating students.

There are other reasons for questioning whether institutional comparisons are likely to lead to meaningful improvements in teaching and learning, as Benjamin asserts. First, the college outcomes assessed by standardized tests represent a very small slice of what is important in education and almost certainly in graduates' postcollege lives. In the studies of standardized tests conducted at UTK, researchers found that the tests measured at most 30% of the institution's general education goals—and those tests included measures of English, mathematics, and social sciences, in addition to measures of critical thinking and writing (Banta & Pike, 1989; Pike, 1988, 1989a, 1989b). In addition, the measures of critical thinking in the CAAP, CLA, and PP may represent different aspects of critical thinking, as Benjamin explains in his third footnote. Benjamin is correct; the fact that a test does not measure all of what is important in education does not automatically invalidate the measure. However, it does raise questions about the appropriateness of making inferences about the quality of education at an institution based on comparisons with other institutions when standardized tests measure a fraction of what is important and measures of supposedly the same construct, such as critical thinking, do not represent the same things.

Second, institutions should question the wisdom of making inferences about institutional performance when research finds few if any statistically significant differences among institutional means and when there is substantially greater variability within institutions than between institutions. Benjamin acknowledges this point, by calling attention to similar circumstances with the National Survey of Student Engagement (NSSE) and other assessments. Just because similar situations exist with other assessments does not mean that it is appropriate to compare institutions. Indeed, Kuh (2007) cautioned against ranking institutions and making simplistic comparisons among institutions using NSSE benchmark scores.

Benjamin's baseball analogy suggests that within-group variability may not be a major concern. He notes that differences among team batting averages are much smaller than the variation in batting averages among players. Our sense is that most baseball fans do not think team batting averages are terribly important—that's why teams play the game and fans watch. Take, for example, the Kansas City Royals, which at the All-Star break in July 2011 had the third highest batting average in the American League, yet they had the worst won-lost record of any team in that league. But there is another problem with the team batting average analogy. Team batting averages are based on all of the at-bats of every player on the team. Institutional means on the standardized assessments Benjamin describes are based on what are in many cases subsamples of students attending an institution that are small and even biased in terms of the disciplines

represented. Given the sampling issues identified previously, as well as the inherent dangers of generalizing from small proportions of the population, within-group variability is a serious concern when making inferences about institutional quality and effectiveness.

There is a third problem with making institutional comparisons using standardized tests. Research on standardized tests has consistently shown that the tests are better measures of individual differences than of educational quality. In the UTK studies cited above, researchers found that as much as 60% of the variation in the scores of individuals was attributable to student characteristics, not educational experiences (see also Pike, 1992). When institutions are the unit of analysis, the correlations between CLA scores and measures of entering ability are extremely high (0.73 to 0.88 for Analytic Writing and 0.78 to 0.92 for performance/critical thinking tasks) (Council for Aid to Education, n.d.). Steedle (2010) reported that the correlation between institutional SAT and CLA scores in his study was 0.93. Given the strength of these relationships, it would appear that what is being measured is the entering abilities and prior learning experiences of students at an institution, and comparisons of institutions based on institutional means would yield little information that is not already available from Barron's Selectivity Index and *U.S. News and World Report* rankings based on ACT/SAT scores. Benjamin and others have argued that these problems may be overcome using measures of value added; others are not convinced. Braun and Wainer (2007, p. 889) concluded, "Given the complexity of educational settings, we may never be satisfied that value added models can be used to appropriately partition the causal effects of the teacher, the school, and the student on measured changes in standardized test scores."

Unfortunately, a focus on individual differences runs counter to the goal of developing tests to assess instructional—and institutional—effectiveness because the best measure of individual differences is student aptitude or ability, not learning.

It is unlikely that standardized tests can overcome problems of insensitivity to educational effects. Hammock (1989) noted that the procedures used to develop norm-referenced tests tend to maximize the ability of the measure to discriminate among individuals. Unfortunately, a focus on individual differences runs counter to the goal of developing tests to assess instructional—and institutional—effectiveness because the best measure of individual differences is student aptitude or ability, not learning. It is not surprising, therefore, that the UTK studies found little or no relationship between scores on standardized tests of generic skills and measures of student learning and development.

One can also raise questions about whether institutions are the appropriate units of analysis when the goal of assessment is improving student learning. Selecting institutions as the units of analysis in assessments implies a high level of consistency in students' educational experiences. However, Chatman's (2007) research showed that there are substantial disciplinary differences in students' educational experiences. Disciplinary differences likely account for high levels of within-institution variability in standardized test scores. These differences also suggest that disciplines, not institutions, would be the appropriate units of analysis in efforts to improve teaching and learning. Of course, this brings us full circle to our initial point that assessments of generic skills like critical thinking are most appropriate at the discipline level.

Benjamin asks, "If an institution is unable to compare itself against its competitors, how will it know how to improve its approach to teaching and learning?" (p. 8) He also asks how, in the absence of a standardized test or benchmark, faculty can decide if a "percentage (is) good, bad, or indifferent?" (p. 9). Faculty routinely develop their own assignments and assess authentic student performance through internships, senior projects, and electronic portfolios. They can administer their assessments in standard circumstances, scoring them using rubrics on which they

have agreed. The same faculty can set performance standards and decide whether student competence is good, bad, or indifferent. Finally, they can use their own students' performance in one term as a benchmark against which to make future comparisons. Controlling their own assessment process provides motivation for faculty to institute improvements aimed at increasing student performance the next time the assessment takes place. Moreover, findings are available immediately and the cost of purchasing expensive commercial instruments is avoided.

Conclusion

In 2007 both of us were members of the APLU-AASCU task forces that recommended the measures to be used in the Voluntary System of Accountability. We made our arguments against employing the CAAP, the CLA, and the PP at that time (Banta & Pike, 2007), but clearly the call by the Commission on the Future of Higher Education (U.S. Department of Education, 2006) for value-added testing and institutional comparisons was too compelling, as reporting value-added statistics using one of those three tests became a VSA requirement. Now, five years later, that requirement has been modified. We appreciate this NILOA forum stimulated by the Benjamin essay as yet another opportunity to discuss the state of the art of measurement as it pertains to standardized tests of generic knowledge and skills.

Controlling their own assessment process provides motivation for faculty to institute improvements aimed at increasing student performance.

References

Baird, L. L. (1988). Diverse and subtle arts: Assessing the generic outcomes of higher education. In C. Adelman (Ed.), *Performance and judgment: Essays on principles and practice in the assessment of college student learning*. Washington, DC: U. S. Government Printing Office.

Banta, T. W., & Pike, G. R. (1989). Methods for evaluating assessment instruments. *Research in Higher Education*, *30*, 455–470.

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

Braun, H., & Wainer, H. (2007). Value-added modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, Psychometrics, pp. 867–889). Amsterdam, The Netherlands: Elsevier.

Chatman, S. (2007, May). *Institutional versus academic discipline measures of student experience: A matter of relative validity* (Center for Studies in Higher Education Research & Occasional Paper Series, CSHE.8.07). Berkeley, CA: University of California.

Council for Aid to Education. (n.d.). *CLA: Frequently asked technical questions, 2007–08*. New York, NY: Author. Retrieved from http://www.collegiatelearningassessment.org/files/CLA_Technical_FAQs.pdf

Hammock, J. (1989). Criterion measures: Instruction vs. selection research. In J. Folger & J. Harris (Eds.), *Assessment in accreditation*. Atlanta, GA: Southern Association of Colleges and Schools.

Herl, H. E., O'Neal, Jr., H. F., Chung, G. K. W. K., Bianchi, C., Wang, S., Mayer, R., Lee, C. Y., Choi, A., Suen, T., & Tu, A. (1999, March). *Final report for validation of problem-solving measures* (CSE Technical Report 501). Los Angeles, CA: University of California, Los Angeles Center for the Study of Evaluation.

Jankowski, N. A., Ikenberry, S. O, Kinzie, J., Kuh, G. D., Shenoy, G. F., & Baker, G. R. (2012). *Transparency & accountability: An evaluation of the VSA College Portrait pilot*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Kuh, G. D. (2007). *Risky business: Promise and pitfalls of institutional transparency*. *Change: The Magazine of Higher Learning*, *39*(5), 30–35.

Paris, D. C. (2011). *Catalyst for change: The CIC/CLA Consortium*. Washington DC: Council of Independent Colleges.

Pike, G. R. (1988, August). *A comparison of the ACT-COMP exam and the ETS Academic Profile. Performance funding report for the University of Tennessee, Knoxville: 1987–88* (Appendix B). Knoxville, TN: University of Tennessee, Center for Assessment Research and Development.

Pike, G. R. (1989a, August). *Comparison of the ACT-COMP and CAAP exams. Performance funding report for the University of Tennessee, Knoxville: 1988–89* (Appendix II). Unpublished report. Knoxville, TN: University of Tennessee, Center for Assessment Research and Development.

Pike, G. R. (1989b, August). *The performance of black and white students on the ACT-COMP exam: An analysis of differential item functioning using Samejima's graded model*. Unpublished research report. Knoxville, TN: University of Tennessee, Center for Assessment Research and Development.

Pike, G. R. (1990a, August). *Comparison of the ACT-COMP exam and the College BASE. Performance funding report for the University of Tennessee, Knoxville: 1989–90* (Appendix I). Unpublished report. Knoxville, TN: University of Tennessee, Center for Assessment Research and Development.

Pike, G. R. (1990b, March). *The performance of males and females on the ACT-COMP exam: An analysis of differential item functioning using Samejima's graded model*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Pike, G. R. (1992). The components of construct validity: A comparison of two measures of general education outcomes. *Journal of General Education*, *41*, 130–159.

Pike, G. R. (2001). Assessment measures: The CRESST problem solving measures. *Assessment Update: Progress, Trends, and Practices in Higher Education*, *13*(4), 14–15.

Renz, D. (2012, in press). Community college strategies. Student learning assessment: A program-level model. *Assessment Update*, *24*(5).

Speary, P. (2002). Community college strategies. The Butler County Community College individualized student assessment pilot project. *Assessment Update*, *14*(3), 8–9, 11.

Steedle, J. T. (2010, April). *Incentives, motivation, and performance on a low-stakes test of college learning*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Sundre, D. L. (2009). Guest editor's note. *Journal of General Education*, *58*(3), vii–ix.

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

Three Ruminations on Seven Red Herrings

Gordon Davies

Roger Benjamin presents seven familiar objections to standardized assessment of student learning in higher education, and refutes them one at a time. He kindly calls them "Seven Red Herrings," not "Seven Deadly Sins." The objections are fair and Roger's refutations are thorough, as we should expect from a leader who has been so deeply engaged with student assessment for so many years. Roger's discussion of assessment is complex and perhaps too "expert" for readers who, like me, are not very familiar with testing theory and practice. But some essays are written by and for experts.

Here are three ruminations on Roger's seven red herrings.

1. Institutional inertia is due to investment in the status quo.

Roger gradually builds to his conclusion, where he says, "The main reason for the relatively little progress that we have achieved in assessment in higher education is institutional inertia." That's it: seven red herrings in a nutshell. "All organizations, including universities and colleges," he goes on, "have set up protocols and decision rules to undertake certain services deemed important for public or private reasons. Institutions, like the individuals that inhabit them, tend to continue their familiar behavior patterns and to resist developing new practices because change requires decisions, and decisions involve risk."

Exactly. The values of higher education institutions have not changed, and neither have the rewards for the "individuals that inhabit them." Despite rhetoric to the contrary, colleges and universities have a huge investment in the status quo, and they are not likely to support changes that may be needed in what they do and how they do it.

2. Our higher education system is driven by prestige.

The more selective an institution's admissions process, and the more research dollars it acquires, the "better" it is. Our national rankings of colleges and universities are based almost entirely on such factors. Their presidents and senior officers are paid -- often handsomely -- to make their institutions elite or more elite. The rewards are substantial for those who make their institutions prestigious, with little or no regard for what ordinary people need from higher education. Chief executives who do not succeed in building or maintaining prestige are replaced. The entire system works to do what is good for institutions, not for people.

Of course, there are exceptions. In his inaugural address more than a decade ago, Lee Todd, the new president of the University of Kentucky, said, "If we make this a nationally prominent research institution and the children of our state do not have better lives, we shall have failed." But his view of the university's role is an exception. Most colleges and universities are concerned with the accepted standards by which excellence is measured. They are not inclined to assess their contributions to the well-being of children. Neither are they anxious to endorse the assessment of teaching and learning, because the rankings of elite institutions—and the compensation that goes with high rankings—do not use such assessments as a measure.

Three ruminations about Benjamin's seven red herrings:

1. *Institutional inertia is due to investment in the status quo.*
2. *Our higher education system is driven by prestige.*
3. *The central mission of American higher education needs to be changed.*

In this frenzied rush for elite status and prestige, no one really wants to know what students are learning or how learning compares across institutions. Roger quite correctly calls for "standardized assessments to permit faculty and administrators to signal how well they are doing in comparison with other higher education institutions." He goes on, "Most importantly, we need good standardized assessment instruments to encourage the development of assessment strategies that directly help faculty to improve teaching and learning in a systematic and continuous manner." But who wants these instruments and strategies? Certainly not the colleges or universities themselves. Reflection and self-evaluation are remarkably absent from these intellectual foundations of our society and its cultures.

About 25 years ago, the Virginia higher education coordinating agency asked institutions to assess student learning, each in its own way, and to report their findings. Several institutions simply refused, until the governor told them he would not support any requests from uncooperative institutions for budget amendments in an upcoming legislative session. All institutions complied, and some results were both humorous and instructive. The sociology department of one college almost fell apart over disagreement about its purpose and how to assess it; the faculty had never considered such a question. Another university chose to interview 100 fourth-year students to get a sense of what they considered to have been valuable to their undergraduate experience. The students praised the beauty of the grounds, the athletic programs, the social life, and a variety of other factors. None mentioned the curriculum or the faculty. "Our complacency has been disturbed," the administration reported. The uncomfortable but simple truth is that colleges and universities respond to a market that measures prestige and elite status. They do not reflect on the basic reasons for their existence and are not committed primarily to meeting the needs of the people whom they nominally serve. And they do not want public disclosure of teaching and learning assessments, especially in comparison with other institutions.

The uncomfortable but simple truth is that colleges and universities respond to a market that measures prestige and elite status. They do not reflect on the basic reasons for their existence and are not committed primarily to meeting the needs of the people whom they nominally serve.

I disagree with only one of Roger's recommendations: "The testing organization should report assessment results for the institutions it tests to those institutions only." This is a dead end. Institutions will use the results selectively, if at all. A comprehensive assessment of teaching and learning will not occur this way.

The current efforts to increase productivity and to enroll and graduate more students from programs of acceptable quality, despite the good intentions behind these efforts, will result in little change. Paying for performance sounds good, but there never has been a funding formula that could not be "gamed," sometimes humorously, sometimes disgracefully, but always to avoid unwanted change. It is easy, and frightening, to imagine how a "pay-for-results" formula could be manipulated.

Institutions always will act in their own best interests. American colleges and universities, both public and private, have embraced a set of values incompatible with the needs of a large portion of the nation's population or with the public good. This leads to my third rumination.

3. The central mission of American higher education needs to be changed.

Most important, we need to reemphasize the public service mission of higher education. Higher education institutions are not just corporations seeking market dominance. For a thousand years—far longer than the lifetime of any private or state-run corporation—colleges and universities have served the public interest. That service is too important to lose.

We certainly want to create one or more fair and reasonable ways to assess student learning, both specific to students' areas of specialization (majors or the equivalent) for use nationally across all institutions of higher education. We probably want to introduce performance assessment that is based on meeting the needs of the people and communities that institutions are intended to serve.

We should reinvigorate the oversight and coordination functions of state higher education agencies. The model for state coordination is more than half a century old. One reason why so many state agencies, governing and coordinating, seem to be held in low esteem may be because they are no longer relevant to the circumstances in which we now live. Perhaps the new model of state coordination should emphasize productivity and assessment: Are institutions graduating enough students from the populations they are supposed to serve, and have the graduates learned to be productive and valuable members of society? This could be a major criterion for a new ranking of colleges and universities.

Perhaps the new model of state coordination should emphasize productivity and assessment: Are institutions graduating enough students from the populations they are supposed to serve, and have the graduates learned to be productive and valuable members of society?

NILOA National Advisory Panel

Joseph Alutto
Provost
The Ohio State University

Trudy W. Banta
Professor
Indiana University-Purdue University Indianapolis

Wallace Boston
President and CEO
American Public University System

Molly Corbett Broad
President
American Council on Education

Judith Eaton
President
Council for Higher Education Accreditation

Richard Ekman
President
Council of Independent Colleges

Mildred Garcia
President
California State University - Fullerton

Susan Johnston
Executive Vice President
Association of Governing Boards

Stephen Jordan
President
Metropolitan State University - Denver

Mary Kalantzis
Dean, College of Education
University of Illinois Urbana-Champaign

Paul Lingenfelter
President
State Higher Education Executive Officers

George Mehaffy
Vice President
Academic Leadership and Change
American Association of State Colleges and Universities

Charlene Nunley
Program Director
Doctoral Program in Community College Policy and Administration
University of Maryland University College

Kent Phillippe
Associate Vice President, Research and Student Success
American Association of Community Colleges

Randy Swing
Executive Director
Association for Institutional Research

Carol Geary Schneider
President
Association of American Colleges and Universities

Michael Tanner
Chief Academic Officer/Vice President
Association of Public and Land-grant Universities

Belle Wheelan
President
Southern Association of Colleges and Schools

Ralph Wolff
President
Western Association of Schools and Colleges

Ex-Officio Members

Timothy Reese Cain
Assistant Professor
University of Illinois Urbana-Champaign

Peter Ewell
Vice President
National Center for Higher Education Management Systems

Stanley Ikenberry
President Emeritus and Regent Professor
University of Illinois

George Kuh
Director, National Institute for Learning Outcomes Assessment
Adjunct Professor, University of Illinois Urbana-Champaign
Chancellor's Professor Emeritus, Indiana University

NILOA Mission

NILOA's primary objective is to discover and disseminate ways that academic programs and institutions can productively use assessment data internally to inform and strengthen undergraduate education, and externally to communicate with policy makers, families and other stakeholders.

NILOA Occasional Paper Series

NILOA Occasional Papers are commissioned to examine contemporary issues that will inform the academic community of the current state-of-the art of assessing learning outcomes in American higher education. The authors are asked to write for a general audience in order to provide comprehensive, accurate information about how institutions and other organizations can become more proficient at assessing and reporting student learning outcomes for the purposes of improving student learning and responsibly fulfilling expectations for transparency and accountability to policy makers and other external audiences.

Comments and questions about this paper should be sent to njankow2@illinois.edu.

About NILOA

- The National Institute for Learning Outcomes Assessment (NILOA) was established in December 2008.
- NILOA is co-located at the University of Illinois and Indiana University.
- The NILOA website went live on February 11, 2009.
- The NILOA research team has scanned institutional websites, surveyed chief academic officers, and commissioned a series of occasional papers.
- One of the co-principal NILOA investigators, George Kuh, founded the National Survey for Student Engagement (NSSE).
- The other co-principal investigator for NILOA, Stanley Ikenberry, was president of the University of Illinois from 1979 to 1995 and of the American Council of Education from 1996 to 2001.
- Peter Ewell joined NILOA as a senior scholar in November 2009.

NILOA Staff

NATIONAL INSTITUTE FOR LEARNING OUTCOMES ASSESSMENT

Stanley Ikenberry, *Co-Principal Investigator*

George Kuh, *Co-Principal Investigator and Director*

Peter Ewell, *Senior Scholar*

Jillian Kinzie, *Associate Research Scientist*

Pat Hutchings, *Senior Scholar*

Timothy Reese Cain, *Senior Scholar*

Natasha Jankowski, *Project Manager and Research Analyst*

Staci Provezis, *Research Associate*

Gianina Baker, *Research Analyst*

Nora Gannon-Slater, *Research Analyst*

Paul Myers, *Research Analyst*

T. Jameson Brewer, *Research Analyst*

Robert Dumas, *Research Analyst*

NILOA Sponsors

Lumina Foundation for Education

The Teagle Foundation

University of Illinois, College of Education

Produced by Creative Services | Public Affairs at the University of Illinois for NILOA. 10.032

Comments on the Commentaries about 'Seven Red Herrings'

Oct 16, 2012 8:30 am by Roger Benjamin
http://illinois.edu/blog/view/915/80826

I am pleased to accept the invitation to briefly respond to some of the points made by those who commented on my "Seven Red Herrings" paper which appeared in the September 2012 issue of the NILOA monthly newsletter. In his Foreword, Peter Ewell predicted that the merits and role of standardized testing will almost certainly continue to be debated. With this in mind, I also offer a few thoughts about what to expect in the future.

Trudy Banta, Gary Pike, and Terrel Rhodes view the promise and potential of standardized testing differently than Margaret Miller and Gordon Davies. Miller sees standardized measures as essential, because the field demands highly reliable and valid assessment tools. At the same time, she believes formative assessment is important as well, albeit for different purposes. Davies goes a step further by saying that colleges and universities must use standardized student learning outcomes measures to assure the public of that these institutions continue to make meaningful, valued contributions both to individuals and the larger society.

Banta and Pike represent the formative end of the assessment continuum. Most of the arguments they presented in their commentary about standardized assessment measures, particularly the Collegiate Learning Assessment (CLA), have appeared previously. Many of their points have been addressed by CLA staff, the Educational Testing Service (ETS), and other researchers, including a summary of approximately 90 studies (Benjamin, et al. 2012). Although my paper was not about the CLA per se, it is worth summarizing several cogent responses available elsewhere to the Banta and Pike's main arguments.

For example, average CLA value-added scores are highly reliable especially at the institution level (freshmen=.94; seniors=.86). Aggregate student motivation is not a significant predictor of aggregate CLA performance, and does not invalidate the comparison of colleges based upon CLA scores. Moreover, the types of incentives that students seem to prefer are not related to motivation and performance.

Although we continue to believe that a no-stakes approach is appropriate for the value-added model in higher education, motivation is a problem for individual student results. CAE (Council for Aid to Education) now offers a version of the CLA protocol, CLA+, which is reliable and valid for individual student performance, as does the Education Testing Service with its Proficiency Profile, and the American College Testing Program with its Collegiate Assessment of Academic Progress. It may well be appropriate in the future to attach stakes to the CLA, which, in turn, likely will increase student motivation to do well.

There is no interaction between CLA task content and field of study. Our researchers find that the CLA protocol measures 30% of the knowledge and skills faculty desire. Results are improved significantly if a representative sample is drawn. Finally, that the CLA is highly correlated with the SAT does not mean the two tests measure the same thing. High school grades combined with the CLA predict freshmen and senior GPA at about the same level as the SAT alone. High school grades plus the SAT and CLA generate a higher prediction than either test alone. This would not be

true if the SAT and CLA measured the same thing.

Banta and Pike are correct in advocating a focus on disciplines, but stray off track by rejecting that standardized test can accurately measure generic cognitive skills (Benjamin et al. 2012). The mean size effect of the growth in student learning outcomes for all colleges testing annually for the past eight years is approximately .73 standard deviations, demonstrating that college attendance is associated with improving these skills.

Banta and Pike suggest there is qualitative evidence to buttress their claims. It would be helpful to know the evidence to which they refer. Measurement scientists privilege statistical-based evidence. This makes conversation between the two groups difficult. Elsewhere I (Benjamin, 2012) explained what I call the assumption of the equality of fields of inquiry. Faculty members are reluctant to question the legitimacy of fields of inquiry that they may not be familiar with. There are solid reasons for this assumption. For example, an obscure field of molecular biology in veterinary medicine focusing on retroviruses in monkeys was critical in helping researchers develop treatments for AIDS. Breakthroughs in one scientific field may lead to startling breakthroughs in others. Measurement science is a field of inquiry that is too well established to be dismissed by colleagues arguing for formative assessments only. For example, Banta and Pike and Rhodes make good arguments for using e-portfolios to assess student learning. However, e-portfolios do not yet pass muster as tools that are sufficiently reliable and valid to obviate the need for appropriate standardized tests for decisions with stakes attached.

Both Davies and Miller want testing organizations to make public student outcome test results. What I should have said was that external demands will require institutions to make their student learning outcomes transparent and that peer review principles aligned with core values of the academy will provide foundational support for higher education leaders creating assessment reporting systems

Peter Ewell noted that faculty prefer to keep assessment results confidential, for internal use only. It is worth noting that testing organizations can achieve greater economies of scale in test development which lowers the price of individual assessments. Aided by recent developments in education technology, there appears to be a burst of innovation in creating assessments for direct use by faculty as instructional tools. Finally, samples of students tested at individual institutions are seldom large enough for the results to be considered sufficiently reliable. More widely used standardized assessments can boost confidence in the results found at individual institutions.

What We Can Expect

The competency-based model now gaining considerable traction will require assessments that corroborate the efficacy of the student learning claimed. Many of those assessments will be standardized tests. There is and will continue to be ample room for formative and standardized tests in postsecondary education. The issue is how to better leverage the virtues of both, for the benefit of improved teaching and learning for the larger societal goals Davies posited.

This, then, is not the time to defend the status quo. Many colleagues may be comfortable in defending positions that marginalize assessment in postsecondary education. Because increasing numbers of private and public leaders believe human capital is the nation's principal resource, debates about how to improve education will continue to grow. The rise of Internet-based education and concerns for the

quality of higher education provided by more traditional means are fueling external demands for increased transparency, restructuring, and accountability.

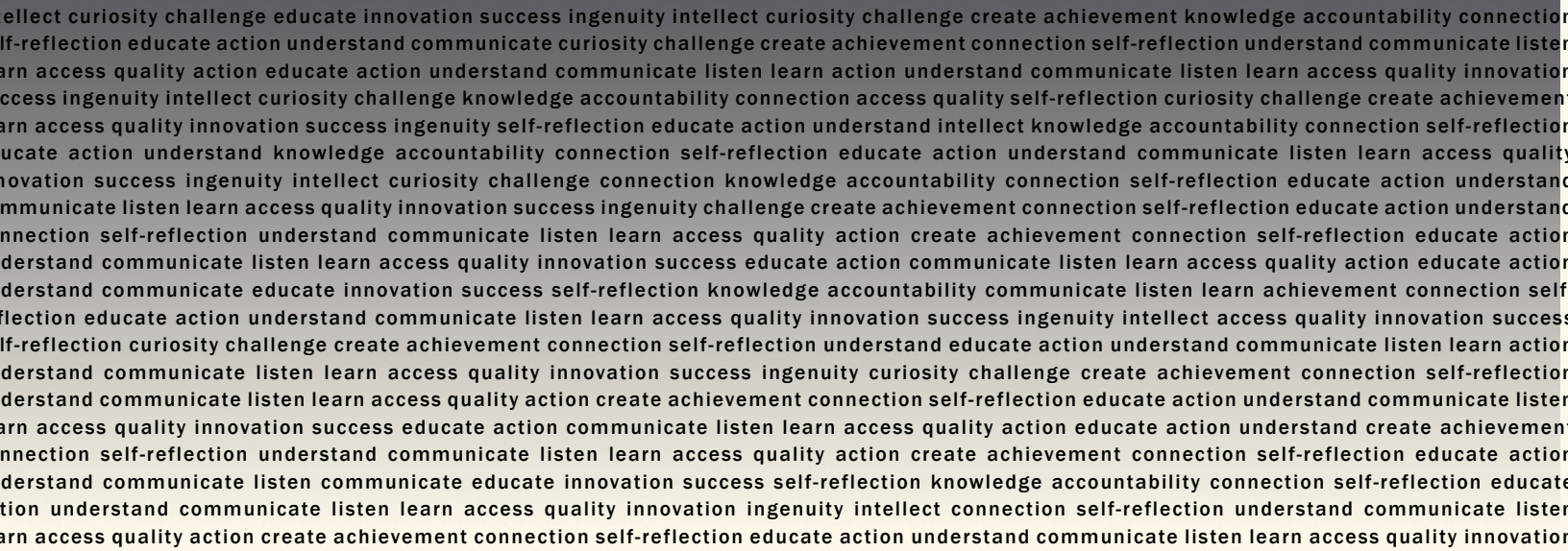
External demands for benchmarking student learning outcomes are destined to increase. However, higher education institutions possess a high level of legitimacy and relative autonomy anchored by department-based governance. The initial challenges for increased transparency of student learning outcomes will come from external forces. Responses to these demands will be developed by innovators within the higher education community. We need all hands on deck to experiment with ways to improve teaching and learning.

Finally, higher education institutions must respond to persistent external demands for more systematic evidence about student learning outcomes. In doing so, the enterprise must also maintain faculty autonomy in determining appropriate assessment approaches; reject college and university ranking systems; privilege efforts to improve student learning; develop assessment protocols that combine standardized and formative assessments; and adhere to peer review principles when constructing accountability systems. About this last observation, there seems little to debate.

References

Benjamin, R. (2012). The new limits of education policy: Avoiding a tragedy of the commons. London: Edward Elgar.

Benjamin, R. Elliot S., Klein S., Steedle, J., Zahner, D., & Patterson, J. (2012). The case for generic skills and performance assessment in the United States and international settings. New York: Council for Advancement of Education.



National Institute for Learning Outcomes Assessment

For more information, please contact:

National Institute for Learning Outcomes Assessment (NILOA)
University of Illinois at Urbana-Champaign
340 Education Building
Champaign, IL 61820

learningoutcomesassessment.org
njankow2@illinois.edu
Fax: 217.244.3378
Phone: 217.244.2155