Research & Occasional Paper Series: CSHE.3.12

CSHE | Center for Studies in Higher Education

UNIVERSITY OF CALIFORNIA, BERKELEY

http://cshe.berkeley.edu/

SERU Project and Consortium Research Paper*

SEARCHING FOR THE HOLY GRAIL OF LEARNING OUTCOMES**

February 2012

John Aubrey Douglass, Gregg Thomson, and Chun-Mei Zhao
UC Berkeley

ABSTRACT

The search for the Holy Grail to measure learning gains started in the US, but the Organisation for Economic Co-operation and Development (OECD) wants to take it global. Here we tell a bit of this story and raise serious questions regarding the validity of the Collegiate Learning Assessment test and suggest there are alternatives. The merit of the CLA as a true assessment of learning outcomes is, we dare say, debatable. In part, the arrival and success of the CLA is a story of markets. In essence, it is a successfully marketed product that is fulfilling a growing demand with few recognized competitors. As a result, the CLA is winning the "learning outcomes race," essentially becoming the "gold standard" in the US. We worry that the CLA's early market success is potentially thwarting the development of other valuable and more nuanced alternatives – whether it be other types of standardized tests that attest to measuring the learning curve of students, or other approaches such as student portfolios, contextually designed surveys on the student experience, and alumni feedback. In a new study published in the journal *Higher Education*, we examine the relative merits of student experience surveys in gauging learning outcomes by analyzing results from the data from the Student Experience in the Research University (SERU) Survey. This essay discusses some of the main points from that article. There are real problems with student self-assessments. But as we argue here, universities can probably learn more about learning outcomes in a wide range of disciplines via properly designed census surveys than by standardized tests like the CLA. At present, we suggest there is tension between the accountability desires of governments and the needs of individual universities who must focus on institutional self-improvement. One might hope that they would be synonymous. But how to make ministries and other policymakers more fully understand the perils of a silver bullet test tool?

It's a clarion call. Ministries of education along with critics of higher education institutions want real proof of student "learning outcomes" that can help justify large national investments in their colleges and universities. How else to construct accountability regimes with real teeth? But where to find the one-size-fits-all test?

In the US, there is a vehicle that claims it can do this – the Collegiate Learning Assessment (CLA) test. In its present form, the CLA is given to a relatively small sample group of students within an institution to supposedly "assess their abilities to think critically, reason analytically, solve problems and communicate clearly and cogently." The aggregated and statistically derived results are then used as a means to judge the institution's overall added value. In the words of the CLA's creators, the resulting data can then "assist faculty, department chairs, school administrators and others interested in programmatic change to improve teaching and learning, particularly with respect to strengthening higher order skills." But can it really do this?

The merit of the CLA as a true assessment of learning outcomes is, we dare say, debatable. In part, the arrival and success of the CLA is a story of markets. In essence, it is a successfully marketed product that is fulfilling a growing demand with few

* The Student Experience in the Research University (SERU) Project and Consortium is a collaborative of major research universities in the US and internationally based at the Center for Studies in Higher Education at UC Berkeley and including the administration of the SERU survey of undergraduates. For more information, see the SERU website at: http://cshe.berkeley.edu/research/seru/

** Adopted from the article Douglass, J.A., Thomson, G., Zhao, C. "The Learning Outcomes Race: the Value of Self-Reported Gains in Large Research Universities," *Higher Education*, February 2012: http://www.springerlink.com/content/hhwt366w82615354/

recognized competitors. As a result, the CLA is winning the "learning outcomes race," essentially becoming the "gold standard" in the US.

But we worry that the CLA's early success is potentially thwarting the development of other valuable and more nuanced alternatives – whether it be other types of standardized tests that attest to measuring the learning curve of students, or other approaches such as student portfolios, contextually designed surveys on student experience, and alumni feedback.

The search for the Holy Grail to measure learning gains started in the US, but the Organisation for Economic Co-operation and Development (OECD) wants to take it global. Here we tell a bit of this story and raise serious questions regarding the validity of the CLA, this global quest, and suggest there are alternatives.

The OECD Enters the Market

In 2008, the OECD began a process to assess if it might develop a test for use internationally. A project emerged: the Assessment of Higher Education Learning Outcomes (AHELO) program would assess the feasibility of capturing learning outcomes valid across cultures and languages, and in part informed by the OECD's success in developing the Programme for International Student Assessment (PISA) – a widely accepted survey of the knowledge and skills essential of students near the end of the compulsory education years.

The proclaimed objective of the AHELO on-going feasibility study is to determine whether an international assessment is "scientifically and practically possible."[1] To make this determination, the organizers developed a number of so-called study "strands." One of the most important is the "Generic Strand," which depends on the administration of a version of the CLA to gauge "generic skills" and competences of students at the beginning and close to the end of a bachelor's degree program. This includes the desire to measure a student's progression in "critical thinking, the ability to generate fresh ideas, and the practical application of theory," along with "ease in written communication, leadership ability, and the ability to work in a group, etc." OECD leaders claim the resulting data will be a tool for the following purposes:

- Universities will be able to assess and improve their teaching.
- Students will be able to make better choices in selecting institutions – assuming that the results are somehow made available publicly.
- Policy-makers will be assured that the considerable amounts spent on higher education are spent well.
- Employers will know better if the skills of the graduates entering the job market match their needs.

Between 10,000 and 30,000 students in more than 16 countries take part in the administration of the OECD's version of the CLA. Full administration at approximately 10 universities in each country is scheduled for 2011 through December 2012.

AHELO's project leaders admit the complexity of developing learning outcome measures, for example, how to account for cultural differences and the circumstances of students and their institutions? "The factors affecting higher education are woven so tightly together that they must first be teased apart before an accurate assessment can be made," notes one AHELO publication.

By March 2010, and at a cost of €150,000 each, the ministries of education in Finland, Korea, Kuwait, Mexico, Norway and the United States agreed to commit a number of their universities to participate in the Generic Strand (i.e. the OECD version of the CLA) of the feasibility study. The State Higher Education Executive Officers – an American association of the directors of higher education coordinating and governing boards – is helping to coordinate the effort in the US. Four states have agreed to participate, including Connecticut, Massachusetts, Pennsylvania, and Missouri. A number of campuses of the Pennsylvania State University agreed to participate in the OECD's version of the CLA with the goal of a spring 2012 administration.

However, the validity and value of CLA is very much in question and the debate over how to measure learning outcomes remains contentious. Many institutions, including most major US research universities, view with skepticism the methodology used by the CLA and its practical applications in what are large institutions, home to a great variety of disciplinary traditions.

The Validity of the CLA?

A product of the Council for Aid for Education (CAE), the CLA is a written test that focuses on critical thinking, analytic reasoning, written communication, and problem solving administered to small random samples of students, who write essays and memoranda in response to test material they have not previously seen. The CAE is technically a non-profit, but has a financial stake in promoting the CLA, much like the Educational Testing Service hawks the SAT.

In the US, the standard administration of CLA involves a cross-sectional sample of approximately 100 first-year students and another 100 fourth-year seniors. It is necessary to keep the sample size small because scoring the narrative is labor intensive. With such a small sample size, there is no guarantee that a longitudinal approach in which the same students are tested will yield enough responses.

CLA proponents justify the cross-sectional approach because students in US colleges and universities often transfer or do not graduate in a four-year period. The cross-sectional design also has the convenience that results can be generated relatively quickly, without having to wait for a cohort to matriculate to their senior year.

Test results derived from these samples are used to represent an institution-wide measure of a university or college's contribution (or value-added) to the development of its students' generic cognitive competencies. Based on these results, institutions can then be compared with one another on the basis of their relative value-added performance.

Proponents of the CLA test claim its value based on three principles:

- First, for accountability purposes, valid assessment of learning outcomes for students at an institution is only possible by rigorously controlling for the characteristics of those students at matriculation.

- Second, by using SAT scores as the control for initial student characteristics, it is possible to calculate the value-added performance of the institution, which is a statistically derived score indicating how the institution fares against what it is expected in terms of student learning. This is done by comparing two value-added scores: one is the actual score, which is the existent difference between freshman and senior CLA test performance; and the other is the predicted score, which is the statistically yielded freshman and senior difference based on student characteristics at entry.

- Third, this relative performance, i.e., the discrepancy between the actual and predicted value-added scores, can in turn be compared to the relative performance achieved at other institutions. Hence the CLA test has accomplished a critical feat in the learning outcomes pursuit: it produces a statistically derived score that is simple and "objective" and that can be used to compare and even rank institutions on how well a college is performing in terms of student learning.

Prominent higher education researchers have challenged the validity of the CLA test on a number of grounds. For one, the CLA and the SAT are so highly correlated. The amount of variance in student learning outcomes after controlling for SAT scores is incredibly small. Most institutions' value-added will simply be in the expected range and indistinguishable from each other. Hence, why bother with the CLA.

The CLA results are also sample-dependent. Specifically, there is a large array of uncontrollable variables related to student motivation to participate in and do well on the test. Students who take CLA are volunteers, and their results have no bearing on their academic careers. How to motivate students to sit through the entire time allotted for essay writing and to take seriously their chore? Some institutions provide extra-credit for taking the test, or provide rewards for its completion. At the same time, self-selection bias may be considerable. On the other hand, there are concerns that institutions may try to game the test by selecting high achievement senior year students. High stakes testing is always subject to gaming. There is no way to avoid institutions cherry-picking – purposefully selecting students who will help drive up learning gain scores.

Other criticisms center on the assumption that the CLA has fashioned a test of agreed-upon generic cognitive skills that is equally relevant to all students. But recent findings suggest that CLA results are, to some extent, discipline-specific. As noted, because of the cost and difficulty of evaluating individual student essays, the design of the CLA relies upon a rather small sample size to make sweeping generalizations about overall institutional effectiveness, it provides very little if any useful information at the level of the major.

To veterans in the higher education research community, the "history lessons" of earlier attempts to rank institutions on the basis of "value-added" measures are particularly telling. There is evidence that all previous attempts at large-scale or campus-wide assessment in higher education on the basis of value-added measures have collapsed, in part due to the observed instability of the measures. In many cases, to compare institutions (or rank institutions) using CLA results merely offers the "appearance of objectivity" that many stakeholders of higher education crave.

The CLA proponents respond by attempting to statistically demonstrate that much of the criticism does not apply to the CLA: for example, regardless of the amount of variance accounted for, the tightly SAT-controlled design does allow for the extraction of valid results regardless of the vagaries of specific samples or student motivation. But ultimately even if the proponents of the CLA are right and their small-sample testing program with appropriate statistical controls could produce a reliable and valid

"value-added" institutional score, the CLA might generate meaningful data in a small liberal arts college, but it appears of very limited practical utility in large and complex universities.

Why? First, the CLA does not pinpoint where exactly a problem lies and which department or which faculty members would be responsible to address the problem. CLA claims that, in addition to providing an institution-wide "value-added" score, it serves as a diagnostic tool designed "to assist faculty in improving teaching and learning, in particular as a means toward strengthening higher order skills."

But for a large, complex research university like the University of California, Berkeley, this is a wishful proposition. Exactly how would the statistically derived result (on the basis of a standard administration of a few hundred freshman and senior test-takers) that, for example, the Berkeley campus was performing more poorly than expected (or relatively more poorly than, say, the Santa Barbara campus in the UC system) assist the Berkeley faculty in improving its teaching and learning?

Second, CLA does not provide enough information on how well a university is doing in promoting learning among students from various backgrounds and life circumstances. This assessment approach is incompatible with the core value of diversity and access championed by the majority of large, public research universities.

Embarking on a "Holy Grail–like" quest for a valid "value-added" measure is, of course, a fundamental value choice. Ironically, the more the CLA enterprise insists that the only thing that really matters for valid accountability in higher education is a statistical test of "value-added" by which universities can be scored and ranked, the more the CLA lacks a broader, "systemic validity," as identified by Henry Braun in 2008:

> Assessment practices and systems of accountability are systemically valid if they generate useful information and constructive responses that support one or more policy goals (Access, Quality, Equity, Efficiency) within an education system without causing undue deterioration with respect to other goals.[2]

"Valid" or not, the one-size-fits-all, narrow standardized test "value-added" program of assessment in higher education promises little in the way of "useful information and constructive responses." A ranking system based on such could only have decidedly pernicious effects, as Cliff Adelman once observed.[3] In Lee Shulman's terms, the CLA is a "high stakes/low yield" strategy where high stakes corrupt the very processes they are intended to support.[4]

For the purposes of institution-wide assessment, especially for large, complex universities, we surmise that the net value of CLA's value-added scheme would be at best unconstructive, and at worst generating inaccurate information used for actual decision-making and rankings.

One Alternative?

In a new study published in the journal *Higher Education*,[5] we examine the relative merits of student experience surveys in gauging learning outcomes by analyzing results from the data from the Student Experience in the Research University (SERU) Consortium and Survey[6] based at the Center for Studies in Higher Education at UC Berkeley.[7] There are real problems with student self-assessments, but there is an opportunity to learn more than what is offered in standardized tests.

Administered since 2002 as a census of all students at the nine undergraduate campuses of the University of California, the SERU survey generates a rich data set on student academic engagement, experience in the major, participation in research, civic and co-curricular activities, time use, and overall satisfaction with the university experience. The survey also provides self-reported gains on multiple learning outcome dimensions by asking students to retrospectively rate their proficiencies when they entered the university and at the time of the survey. SERU results are then integrated with institutional data.

In 2011, the SERU Survey was administered at all nine University of California undergraduate campuses, and to students at an additional nine major research universities in the US, all members of the Association of American Universities (AAU), including the Universities of Michigan, Minnesota, Florida, Texas, Rutgers, Pittsburgh, Oregon, North Carolina and the University of Southern California. (A SERU-International Consortium[8] has recently been formed with six "founding" universities located in China, Brazil, the Netherlands, and South Africa.)

SERU is the only nationally administered survey of first-degree students in the US that is specifically designed to study policy issues facing large research universities. It is also one of four nationally recognized surveys for institutional accountability for research universities participating the Voluntary System of Accountability initiative in the US. The other surveys include the College Student Experiences Questionnaire, the College Senior Survey, and the National Survey of Student Engagement.

The technique of self-reported categorical gains (e.g., "a little", "a lot") typically employed in student surveys has been shown to have dubious validity compared to "direct measures" of student learning. The SERU survey is different. It uses a retrospective posttest design for measuring self-reported learning outcomes that yields more valid data. In our exploration of that data, we show connections between self-reports and student GPA and provide evidence of strong face validity of learning outcomes based on these self-reports.

The overall SERU survey design has many other advantages, especially in large, complex institutional settings. It includes the collection of extensive information on academic engagement as well as a range of demographic and institutional data. The SERU dataset sheds light on both the variety of student backgrounds and the great variety of academic disciplines with their own set of expectations and learning goals.

Without excluding other forms of gauging learning outcomes, we conclude that designed properly, student surveys offer a valuable and more nuanced alternative in understanding and identifying learning outcomes in the university environment.

But we also note the tension between the accountability desires of governments and the needs of individual universities who should focus on institutional self-improvement. One might hope that they would be synonymous. But how to make ministries and other policymakers more fully understand the perils of a silver bullet test tool?

The Lure of the Big Test

Back to the politics of the CLA. This test is a blunt tool, creating questionable data that serves immediate political ends. It seems to ignore how students actually learn and the variety of experiences among different sub-populations. Universities are more like large cosmopolitan cities full of a multitude of learning communities, as opposed to a small village with observable norms. In one test run of the CLA, a major research university in the US received data that showed students actually experienced a decline in their academic knowledge – a negative return? It seems highly unlikely.

But how to counteract the strong desire of government ministries, and international bodies like the OECD, to create broad standardized tests and measures of outcomes? Even with the flaws noted, the political momentum to generate a one-size-fits-all model is powerful. The OECD's gambit has already captured the interest and money of a broad range of national ministries of education and the US Department of Education.

What are the chances the "pilot phase" will actually lead to a conclusion to drop the pursuit of an higher education version of PISA? Creating an international "gold standard" for measuring learning outcomes appears too enticing, too influential, and too lucrative for that to happen – although we obviously cannot predict the future.

It may very well be that data and research offered in our study that uses student survey responses will be viewed as largely irrelevant in the push and pull for market position and political influence. Government's love to rank and this might be one more tool to help encourage institutional differentiation – a goal of many nation-states.

But for universities who desire data for making actionable improvement we argue that student surveys, if properly designed, offer one of the most useful and cost-effective tools. They also offer a means to combat simplistic rankings generated by CLA and similar tests.

REFERENCES/LINKS

[1] See AHELO website: http://www.oecd.org/document/41/0,3343,en_2649_35961291_42295209_1_1_1_1,00.html
[2] http://www.nciea.org/publications/RILS08_HB_092508.pdf
[3] http://www.edweek.org/ew/articles/2006/11/08/11adelman.h26.html
[4] http://www.unr.edu/assess/AssessmentMattersFiles/LeeShulman_Change_07.pdf
[5] http://www.springerlink.com/content/hhwt366w82615354/
[6] http://cshe.berkeley.edu/research/seru/consortium.htm
[7] http://cshe.berkeley.edu/
[8] http://cshe.berkeley.edu/research/seru/intlconsortium.htm