High Educ (2012) 64:317–335
DOI 10.1007/s10734-011-9496-x

The learning outcomes race: the value of self-reported gains in large research universities

John Aubrey Douglass · Gregg Thomson · Chun-Mei Zhao

Published online: 4 February 2012

Abstract Throughout the world, measuring "learning outcomes" is viewed by many stakeholders as a relatively new method to judge the "value added" of colleges and universities. The potential to accurately measure learning gains is also a diagnostic tool for institutional self-improvement. This essay discussed the marketisation of learning outcomes tests, and the relative merits of student experience surveys in gauging learning outcomes by analyzing results from the University of California's Undergraduate Experience Survey (Student Experience in the Research University Survey: SERU-S). The SERU-S includes responses by seniors who entered as freshmen on six educational outcomes self-reports: analytical and critical thinking skills, writing skills, reading and comprehension skills, oral presentation skills, quantitative skills, and skills in a particular field of study. Although self-reported gains are sometimes regarded as having dubious validity compared to so-called "direct measures" of student learning, the analysis of this study reveals the SERU survey design has many advantages, especially in large, complex institutional settings. 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 broad tapestry of higher education institutions. We discuss the politics of the learning outcomes race, the validity of standardized tests like the Collegiate Learning Assessment (CLA), and what we can learn from student surveys like SERU-S. We also suggest there is a tension between what meets the accountability desires of governments and the needs of individual universities focused on self-improvement.

Keywords Learning outcomes ▫ Standardized tests ▫ Global higher education markets ▫ AHELO ▫ Student academic engagement

J. A. Douglass (✉) ▫ G. Thomson ▫ C.-M. Zhao
Center for Studies in Higher Education, University of California-Berkeley, Berkeley, CA, USA
e-mail: douglass@berkeley.edu

G. Thomson
e-mail: gthomson@berkeley.edu

C.-M. Zhao
e-mail: mzhao@berkeley.edu

In the US and throughout the world, interest in measuring learning outcomes at all levels of education has merged over the past decade. In higher education, "learning outcomes" are viewed by many stakeholders, including lawmakers and advocates of new and more expansive *accountability* regimes, as a method to measure the value added, and to a large extent the quality and effectiveness, of colleges and universities.

For universities, collecting some form of learning outcomes data is a growing component of institutional *assessment* and a route for institutional self-improvement. But external forces heavily influence the global search for data and analysis. Government ministries, along with accrediting agencies, the media, and critics of higher education, desire a universal tool to measure learning outcomes at the campus level, and that can be compared across institutions, regions, and perhaps even countries.

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

The merit of the CLA as a valid assessment of learning outcomes is debatable. As we discuss in the following, the arrival and success of the CLA is, in part, a story of markets. In essence, it is a successful product that is fulfilling a growing demand with few recognized competitors. As a result, the CLA's is winning the "learning outcomes race," becoming the "gold standard" in the US and globally. It is also 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 a student portfolios or surveys on the student experience.

This study explores the question of what is most the effective mechanism for gauging learning outcomes: standardized testing, student self-reported gains, discipline based assessments, or some combination? We focus on the validity of the CLA versus that of student self-assessed learning outcomes generated by the Student Experience in the Research University Survey (SERU-S).

Administered since 2002 as a census to all students at the nine undergraduate campuses of the University of California, SERU-S generates a rich data set on student participation in research, civic and co-curricular activities, time use, student self-reported gains at multiple learning outcome dimensions from the beginning of their college career until the time of survey taking place, and student experience in their major and overall satisfaction with their university experience. SERU results are then integrated with institutional data. In addition, the SERU Survey is now administered at an additional nine universities in the US, all who are members of the Association of American Universities (AAU).[1] SERU is the only nationally administered survey of first-degree students in the US that is specifically designed around policy and scholarly issues that face large research-intensive

[1] In 2011 the SERU Consortium included 18 major US research universities—the nine general campuses of the University of California System, plus the Universities of Michigan, Minnesota, Florida, Texas, Rutgers, Pittsburgh, Oregon, North Carolina and the University of Southern California. Fifteen are members of the prestigious American Association of Universities (AAU). For further information on the SERU Consortium, see: http://cshe.berkeley.edu/research/seru/consortium.htm.

universities. It is also identified as one of four nationally recognized surveys for institutional accountability for research universities participating the Voluntary System of Accountability initiative.[2]

Although self-reported gains are sometimes regarded as having dubious validity compared to so-called "direct measures" of student learning, the analysis of this study reveals the SERU survey design has many advantages, especially in large, complex institutional settings. 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 broad tapestry of higher education institutions. In the following, we discuss the politics of the learning outcomes race, the validity of standardized tests like the CLA, and what we can learn from student surveys like SERU-S. We also suggest there is a tension between what meets the accountability desires of governments and the needs of individual universities focused on self-improvement.

The political landscape of learning outcomes movement

The interest in generating a means to measure the learning outcomes of undergraduates in US universities and colleges emerged in the 1990s. But it gained elevated significance in the news media and among state governments under President George W. Bush's administration (2001–2009). In 2005, Secretary of Education Margaret Spellings convened a special commission to focus on how to make higher education institutions more accountable in light of rising public and private investment in American colleges and universities. Reflecting to some degree the structural approach of the "No Child Left Behind" legislation that focused on reform in K-12 education, the "Spellings Commission" advocating the building of a similar learning assessment program in US higher education largely through the use of standardized test. In its final September 2006 report, the commission imagined two routes for greater accountability:

- The development and wide use of some sort of standardized test to measure value added
- New federal guidelines for the nation's network of accrediting bodies to help develop national standards and comparative review of institutional performance

An institution should "gather evidence about how well students in various programs are achieving learning goals across the curriculum and about the ability of its graduates to succeed in a challenging and rapidly changing world," stated the report, "and the information should be used, as it historically has been, to help the institutions figure out how best to improve their performance" (Spellings Commission 2006).[3]

[2] For the student experiences and perceptions category of the VSA, participating institutions are required to report data from one of four surveys: the College Student Experiences Questionnaire, the College Senior Survey, the National Survey of Student Engagement, or the SERU Survey (or what is known in the UC system as the University of California Undergraduate Experience Survey).

[3] The Spellings Commission was established on September 19, 2005, by U.S. Secretary of Education Margaret Spellings. The nineteen-member Commission was charged with recommending a national strategy for reforming post-secondary education, with a particular focus on how well colleges and universities are preparing students for the 21st-century workplace, as well as a secondary focus on how well high schools are preparing the students for post-secondary education. In the report, released on September 26, 2006, the Commission focuses on four key areas: access, affordability (particularly for non-traditional students), the standards of quality in instruction, and the accountability of institutions of higher learning to their constituencies (students, families, taxpayers, and other investors in higher education).

Although without significant authority over state higher education systems, the federal commission heightened an ongoing debate over the ideal of measuring learning outcomes. There was also debate over the appropriate use of such data—for example, as a means for identifying poor institutional performers, for conditioning federal and state funding, and for informing potential students and their families.

The call for added accountability, the emphasis on testing, and the refocusing of the voluntary national accreditation system have had the beneficial effect of increasing the higher education community's attention to more systematically evaluating teaching and learning. On the heels of the Spelling Commission, the National Association of State Universities and Land-Grant Colleges and the American Association of State Colleges and Universities collaborated to create a Voluntary System of Accountability (VSA) that requires participating institutions to report learning outcomes using one of three competing standardized tests of undergraduate "higher order skills." This includes the Assessment of Academic Proficiency test marketed by the ACT Inc, the purveyor of a test that competes with the SAT for testing high school student applying to college; the Measure of Academic Proficiency and Progress offered by the Educational Testing Service, owner of the SAT; and the CLA.

The notion that standardized testing is the appropriate way to assess learning outcomes at the university level has not been universally accepted, however. In fact, in 2007 the University of California explicitly rejected this component of the VSA, noting that "using standardized tests on an institutional level as measures of student learning fails to recognize the diversity, breadth, and depth of discipline-specific knowledge and learning that takes place in colleges and universities today."[4]

In 2008 the Consortium on Financing Higher Education (COFHE) released its statement on assessment in which it firmly rejected standardized testing:

> Based on our experience, we are skeptical about efforts to make this kind of assessment through standardized tests, including those that purport to measure critical reasoning. … [A]ssessment experts are far from agreement about whether "value added" can be measured accurately across diverse institutions. … [W]e do not endorse any approach that depends solely on a single standardized measure or even a single set of standardized measures. (COFHE 2008)

In addition to the COFHE membership of 31 leading private colleges and universities, the statement on assessment was endorsed by dozens of others, including the University of California, Berkeley. Ironically, by early 2008, Secretary Spellings herself apparently no longer held the view that the one-measure-fits-all-institutions approach advocated by the Spellings Commission was appropriate. "All colleges should be allowed to describe their own unique missions," she stated before the National Press Club, "and be judged against that." She went on to say, "That is totally within the jurisdiction of each institution."[5]

Regardless of the opposition to standardized testing to assess learning outcomes, in the US the imperative to measure and report on student learning outcomes for accreditation and public accountability remains strong. A number of critiques of American higher education have also charged that students are not learning much in an era of increased tuition costs. A recent study, *Academically Adrift*, used CLA data to claim that there were virtually no learning gains among traditionally aged students (18–24 year olds) in a variety

[4] UC President Robert C. Dynes quoted in Scott Jaschik, "Accountability System Launched," Inside Higher Education, Nov. 12, 2007.

[5] Speech before the National Press Club, report in *The Chronicle of Higher Education*, Feb. 1, 2008.

of colleges and universities. "How much are students actually learning in contemporary higher education? The answer for many undergraduates, we have concluded, is not much," wrote the authors of the study (Arum and Roska 2011).[6]

Meanwhile, international pressure has been mounting to find some comparative mechanism to gauge the value added effects of higher education institutions (HEI's). Much of the focus in the past has been on the economic benefits of public investment in university generated research, and more generally on the role of universities in creating talented labor pools. Now this interest is expanding to measuring learning outcomes.

In 2008, the OECD began a process to assess if it might develop a test and other assessment tools internationally. The Assessment of Higher Education Learning Outcomes (AHELO) was then established and is currently assessing the feasibility of capturing learning outcomes on an international scale by creating measures that would be valid across cultures and languages. Venturing into the higher education market is in part informed by the OECD's success in developing the Programme for International Student Assessment (PISA)—a widely accepted survey of students near the end of compulsory education intended to assess their knowledge and skills deemed essential for full participation in society.

The proclaimed object of the feasibility study is to determine whether an international assessment is "scientifically and practically possible."[7] To make this determination, there are a number of study "strands." One of the most important strands is the administration of a version of the CLA to gauge "generic skills" and competences of students who are almost at the end of a bachelor's degree program This includes the hope to measure a student's progress in, "critical thinking, the ability to generate fresh ideas, and the practical application of theory. Ease in written communication, leadership ability, and the ability to work in a group, etc. could also be included." OECD leaders claim the resulting data will be a tool for:

- Universities to assess and improve their teaching.
- Students to make better choices in selecting institutions.
- Policy-makers to make sure that the considerable amounts spent on higher education are spent well.
- Employers to know if the skills of the graduates entering the job market match their needs

The OECD planned that between 10,000 and 30,000 higher education students in more than 16 countries will take part 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 that the complexity of

[6] Authors Richard Arum at New York University, and Josipa Roksa at the University of Virginia charged that many undergraduates, are "drifting through college without a clear sense of purpose is readily apparent." Based on CLA data, they tracked the academic gains (or stagnation) of 2,300 students of traditional college age enrolled at a range of 4-year colleges and universities. 45% of students "did not demonstrate any significant improvement in learning" during the first 2 years of college; 36% of students "did not demonstrate any significant improvement in learning" over 4 years of college. Those students who do show improvements tend to show only modest improvements. Students improved on average only 0.18 standard deviations over the first 2 years of college and 0.47 over 4 years. What this means is that a student who entered college in the 50th percentile of students in his or her cohort would move up to the 68th percentile 4 years later—but that's the 68th percentile of a new group of freshmen who haven't experienced any college learning.

[7] See AHLO website: http://www.oecd.org/document/41/0,3343,en_2649_35961291_42295209_1_1_1_1,00.html.

developing learning outcome measures that account for cultural differences and the circumstances of students and their institutions is significant. "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 (AHELO 2011).

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 of the feasibility study. A year earlier, the US Department of Education announced that it would join AHELO, with a focus on the generic skills strand. Four states have agreed to participate, including Connecticut, Massachusetts, Pennsylvania, and Missouri. The State Higher Education Executive Officers (SHEEO)—an association of the directors of higher education coordinating and governing boards—will help coordinate the effort. 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.

Both in the US and around the world, CLA has emerged as one primary tool for gauging generic learning outcomes of college students. However, as noted, 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 research universities, view with skepticism both the methodology used by the CLA, and its practical applications in what are large institutions with strong disciplinary traditions.

The CLA: the unconstructive value of "value added"

A product of the Council for Aid for Education (CAE), the CLA offers a written test that focuses on critical thinking, analytic reasoning, written communication, and problem solving that is administered to small random samples of freshmen in the fall and seniors in the spring. Students write essays and memoranda in response to test material they have not previously seen. Under the auspices of the CAE, the CLA is seemingly modeled on the Educational Testing Service, the extremely successful purveyors of the SAT and similar standardized tests.

The CLA is administered as a cross-sectional sample of approximately 100 first-year students and another 100 seniors (fourth-year) each year. It is necessary to keep the sample size small because the scoring of the narrative is labor intensive. 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 the institution's contribution (or value-added) to the development of the generic cognitive competencies of their students. Based on these results, institutions can then be compared on the basis of their relative value-added performance.

CAE claims the CLA has three purposes:

- 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 (Klein et al. 2005, 2007).
- Second, by using SAT scores as the control for initial student characteristics, it is possible on the basis of small samples to calculate the actual value-added of the

institution (i.e., the difference between freshman and senior test performance) and compare it to the predicted value-added, which is the predicted freshman and senior difference based on student characteristics at entry.

- Third, this relative performance or value-added can in turn be compared to the relative performance or value-added achieved at other institutions, hence providing the most valid or fair comparison of how well a college is performing in terms of student learning (Klein et al. 2007, 2008).

Prominent higher education researchers (Banta 2006, 2007, 2009; Pike 2006, etc.) have questioned the CLA test on a number of grounds. For one, the CLA and the SAT are so highly correlated that the amount of variance in student learning outcomes to be accounted for after controlling for SAT scores is incredibly small and most institutions' value-added will simply be in the expected range.

The results are also sample-dependent in ways not until recently recognized by CLA proponents. Specifically, there is a large array of variables related to student motivation to do well on what is in effect a low-stakes test for students. Students who take CLA are volunteers, and their results have no practical 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 remains a conundrum (Hosch 2010). Some institutions have attempted to provide extra-credit for taking the test, or to provide rewards to its completion. There are also concerns that institutions may try to game the test by selecting high achievement senior year students.

At the same time, a design that compares the test performance of a sample of freshmen and a sample of seniors cannot isolate institutional value-added from other characteristics of institutions and their students that affect student learning, but have nothing directly to do with the instructional quality and effectiveness of an institution.

Other criticisms center on the assumption that the CLA has fashioned tests of agreed-upon generic cognitive skills that are equally relevant to all students (Pike 2006), but recent findings suggest that CLA results are, to some extent, discipline-specific (Arum and Roska 2008). 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. The resulting data is rather erratic as it includes a broad array of fields of studies. CLA results are more dependent on the composition of programs and students than the actual value-added learning of the universities.

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 change measures (Adelman 2006; Banta 2006, 2007; Pike 2006). In many cases, to compare institutions (or rank institutions) using CLA results merely offers the "appearance of objectivity" many stakeholders of higher education crave.

The CAE's response attempts to demonstrate statistically that much of this 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 (Klein et al. 2007, 2008). 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 utility in large and complex universities due to the following reasons.

First, the 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, this is a wishful proposition—exactly how would the statistically derived result (on the basis of a few hundred freshman and senior test-takers) showing that, for example, the Berkeley campus was performing more poorly than expected (or relatively more poorly than, say, the Santa Barbara campus) assist the Berkeley faculty in improving its teaching and learning? It does not pinpoint where exactly the problem lies and which department or which faculty members would be responsible to address the problem. In reality, this news would surely generate "more heat than light" and could offer no guidance whatsoever in terms of institutional self-improvement.

Second, CLA does not provide any information regarding how well a university is doing regarding its 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.

Finally, 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 Braun:

> 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. (Braun 2008)

"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 Adelman (2006) observes. 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 (2007). 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 unconstructive.

The value of self-reported learning outcomes in a census design

The SERU Survey adopts a census design, which makes it possible to provide data down to the level of individual academic program and student subpopulations of interest. This is especially valuable for large-scale research universities, to help meet their comprehensive informational needs.

At the same time, SERU-S offers an approach that sets it apart from conventional undergraduate surveys in terms of assessing learning outcomes. The survey instrument adopts a retrospective pretest and a current posttest (that is, a "then" and "now") design to measuring student learning outcomes rather than the commonly used rating of change (or

the perceived change method) used by a number of other surveys of student experience in the US. Specifically the SERU Survey asks students to rate their level of proficiency at two time points (when they started at the university and now), using a six-point scale (Very Poor, Poor, Fair, Good, Very Good, Excellent) on a series of educational outcomes.

The retrospective posttest design is drawn from the field of program evaluation in the work of Howard and others (Howard 1980; Howard et al. 1979; Howard and Dailey 1979), who challenged the conventional wisdom that the most valid way to measure program effects or gains is to use a pretest–posttest design. Howard identified the "response-shift bias", which means program participants are likely to have a more informed frame of reference as a consequence of their experience in the program, therefore making posttest evaluations of their proficiencies both lower and more accurate than their pretest evaluations.

With this insight, assessment of program or treatment effects do not need to rely as heavily on the more costly pretest–posttest evaluation design and could often substitute the retrospective posttest design. Some research has pointed out that the retrospective posttest design may produce upwardly biased ratings, often due to motivational bias (e.g., social desirability) or systematic cognitive bias such as self-enhancement, implicit theory of change, and effort justification (Hill and Betz 2005; Taylor et al. 2009). On the other hand, further research found that retrospective pretest design produced the least social desirability bias compared to other design approaches (such as the perceived change method) (Lam and Bengo 2003; Krosnick 1991).

Several observations and generalizations emerge from the practice of the retrospective pretest method in program evaluation and other fields. If the goal is to accurately capture how change is experienced subjectively by students by program, this method is especially useful. Where what is being rated is salient to the participants' sense of self, the "then" and "now" method may be more appropriate despite the obvious heightened social desirability bias. A key condition to use the retrospective pretest method is that if the costs for overestimating program effects are not great, the advantages of using this approach can offset the potential biases.

This condition applies to any large-scale effort to assess and report learning outcomes in higher education. More importantly, this approach allows us to capture accurately how different populations of students (e.g., students in different majors) characterize their own learning gains at a large university and under what conditions should contribute considerably to our potential understanding of the complexities of learning outcomes.

Given the political realities of accountability, an institution's entirely transparent though favorably biased presentation of learning gains as reported by students themselves surely has less potential downside than the possibility of coming out on the short end of a perhaps unstable and certainly opaque (for the public) "value-added" ratings scheme such as the CLA.

Granted, student self-reports of learning are only indirect indicators and tend to be upwardly biased. On the other hand, the large-scale census design allows us to amass tremendous amounts of learning and other self-reported educational outcomes data, thereby providing an opportunity to conduct analyses to understand the extent how self-reports are useful in validating and reporting learning outcomes.

The SERU-S census design, therefore, offers a rich set of student retrospective pretest or "then" and "now" self-assessment data. We can examine self-reported educational outcomes across a large number of domains for students at every point of their academic careers, across and within different fields of study and for any number of student

populations. Having retrospective pretest items across so many different content areas also gives us the ability to help assess and control for the tendency to exhibit improvement biases.

Examining the validity of self-reported learning outcomes

In an effort to further explore the validity of SERU data in assessing student learning gains from a few angles, we examined the responses of seniors who entered as freshmen on six of the self-reported educational outcomes on SERU-S 2008: analytical and critical thinking skills, writing skills, reading and comprehension skills, oral presentation skills, quantitative skills, and skills in a particular field of study. Specifically, we examined the relationship between self-reported gains with other student and program characteristics such as University of California Grade Point Average (GPA, the grades given to students in each course on a 4-point scale), gender, race/ethnicity, immigrant generation, and major to understand the meaning the self-reported gains. Omitting respondents with missing data, the study includes about 12,500 sets of responses.

Table 1 shows how UC seniors assess their learning gains in each of the six areas. While seniors are more likely to rate themselves to be more proficient currently than when they began at the university in all six areas, what is noteworthy is how the magnitude of the gains varies across the six areas. There is less gain reported for quantitative skills in particular, which makes sense given that the majority of students major in non-quantitative-based fields. At the other extreme, self-reported gains are highest for knowledge of a specific field of study; that is, an area that cuts across all majors. These results, then, seem to have credible face-validity.

We then examined the relationship of self-reported learning gains with another more direct measure of learning, namely college GPA. The relationship between student self-reports and overall cumulative UC GPA is presented in Table 2.

Overall, higher GPA tends to be connected with higher levels of self-reported gains: this appears to make intuitive sense in that students who achieved higher grades feel like they have learned more. One exception is in Quantitative skills, where higher GPA is linked to lower gains.

The relationships of UC GPA and student self-reports vary by skill area suggests that student assessments using the SERU-S approach have some degree of validity as indicators of learning outcomes. Skill areas are less uniformly related to academic achievement across all majors. Specifically, the relationship is weakest for quantitative skills (gains are actually the lowest for the highest GPA students); conversely, it is strongest for critical and analytical thinking and field of study.

Table 1 Percent rating skills as "Very Good" or "Excellent" across six domains

	Began (%)	Now (%)	Gain (%)
Quantitative skills	28	39	+11
Oral presentation	18	56	+38
Writing clearly	24	62	+38
Reading academic	22	71	+49
Critical thinking	24	76	+52
Field of study	6	76	+70

Table 2 Percent rating skills as "Very Good" or "Excellent" across six domains by current cumulative UC GPA category

	Began (%)	Now (%)	Gain (%)
Quantitative			
Under 2.8	23	35	+12
2.8–3.19	23	36	+13
3.2–3.59	26	36	+10
3.6 & higher	34	41	+6
Oral presentation			
Under 2.8	18	53	+34
2.8–3.19	17	51	+35
3.2–3.59	17	55	+38
3.6 & higher	19	57	+38
Writing			
Under 2.8	19	53	+34
2.8–3.19	20	55	+36
3.2–3.59	23	61	+38
3.6 & higher	29	69	+39
Reading			
Under 2.8	19	59	+39
2.8–3.19	19	62	+43
3.2–3.59	21	70	+49
3.6 & higher	25	77	+52
Critical thinkng			
Under 2.8	19	63	+44
2.8–3.19	20	67	+48
3.2–3.59	23	75	+52
3.6 & higher	29	82	+53
Field of study			
Under 2.8	6	62	+56
2.8–3.19	6	70	+64
3.2–3.59	6	76	+70
3.6 & higher	7	82	+75

Further examination of the relationship between self-reported gains and field of study and students demographics provides another layer of analysis.

As shown in Table 3, they are significantly related, further supporting the validity of the self-reported gains: students who were not born in the US have the highest level of gain in Oral communication and presentation; students from STEM fields reported highest level of gain in Quantitative skills, whereas humanities and social sciences students reported the lowest level of improvement in Quantitative skills. On the other hand, humanities and social sciences students reported highest levels of gains in writing, reading and critical thinking skills.

Looking deeper, the results seem to indicate that a multiplicity of factors may contribute to student self-ratings when using the retrospective pretest method. However, because these factors are to a substantial degree interrelated, we then chose to examine the effects of the factors in combination. To do this, student responses were analyzed using a five factor, $2 \times 2 \times 2 \times 2 \times 2$ design. These factors include:

Table 3 Senior self-reports by ethnicity, demographics and field of study

	Quant	Oral	Writing	Reading	Thinking
Ethnicity					
Began					
Asian	29%	17%	20%	19%	20%
Black	21%	22%	25%	30%	24%
Latino	20%	20%	21%	24%	21%
White	28%	25%	35%	32%	35%
Now					
Asian	39%	46%	47%	56%	60%
Black	33%	65%	70%	78%	82%
Latino	33&	61%	64%	75%	78%
White	39%	56%	71%	79%	84%
Gains					
Asian	+10	+29	+27	+37	+40
Black	+12	+43	+45	+48	+58
Latino	+13	+41	+43	+51	+57
White	+11	+31	+36	+47	+49
Immigration					
Began					
Student not born in US	28%	18%	17%	20%	20%
Parent(s) not born in US	25%	22%	22%	24%	22%
Both parents born in US	26%	27%	34%	34%	36%
Now					
Student not born in US	41%	42%	43%	54%	56%
Parent(s) not born in US	36%	44%	51%	58%	59%
Both parents born in US	36%	50%	66%	72%	78%
Gains					
Student not born in US	+13	+24	+26	+34	+36
Parent(s) not born in US	+11	+22	+29	+34	+37
Both parents born in US	+10	+23	+32	+38	+42
Field of study					
Began					
Engineering, math, science	39%	17%	26%	23%	31%
Biological sciences	34%	19%	25%	23%	25%
Social sciences	22%	22%	25%	26%	26%
Humanities	18%	26%	33%	33%	31%
Now					
Engineering, math, science	74%	52%	44%	59%	63%
Biological sciences	49%	48%	50%	64%	59%
Social sciences	28%	53%	65%	70%	68%
Humanities	14%	54%	75%	77%	73%
Gains					
Engineering, math, science	+35	+35	+18	+36	+32
Biological sciences	+15	+29	+25	+31	+34

Table 3 continued

	Quant	Oral	Writing	Reading	Thinking
Social sciences	+06	+31	+40	+44	+42
Humanities	−03	+28	+42	+44	+42

- UC GPA: <3.2 versus ≥3.2
- Major: Science (STEM disciplines) versus non-science
- Immigration Status: Immigrant (both parents not born in US) versus non-immigrant
- Gender: Male versus female
- Ethnicity: Asian versus non-Asian

This design yields 32 separate combinations or 16 "controlled" comparisons for each of the five factors. For example, in examining the relationship of UC GPA to self-ratings we compared the ratings of male immigrant Asian science respondents in the two GPA categories, the ratings of female immigrant Asian science respondents in the two GPA categories, and so forth. Unweighted averages for the 16 comparisons for each of the five factors across the six skill domains are shown in Table 4.

This analysis suggests that UC GPA, field of study, and ethnicity are all associated with substantial differences in student self-ratings of educational outcomes *even after controlling for other factors*. For example, after controlling for other factors, gains by UC GPA are greater for field of study and reading academic material; for field of study the greater gains by science students in quantitative skills are offset by equally greater gains by non-science students in writing clearly; and for ethnicity, Asian student percentage gains are double-digits less for all areas except quantitative skills. Immigrant generation has modest effects and, with the exception of quantitative skills, there are no differences by gender.

To appreciate the magnitude of the combined effect of UC GPA, field of study, and ethnicity on self-ratings, three-way crosstabs were run for current skill ratings for each of the six domains. As can be seen in Table 5, the joint effects of UC GPA, field of study, and ethnicity on the proficiency ratings of University of California seniors can be quite dramatic. For example, for "writing clearly and effectively" the range is from 43 to 80% rating themselves as "Very Good" or "Excellent". The different relative magnitude of each of the factors across different skill domains in ways that "make sense" is also worth noting, offering support to the validity of the self-reported gains.

Our approach here unveils, yet probably underestimates the full impact of various factors on self-ratings. For example, certain fields of study (e.g., engineering, within science) and more differentiation with UC GPA would yield more extreme differences. The fact that Asian students rate themselves lower even after controlling, at least broadly, for other factors is, of course, very intriguing and may be our first hint of important cultural differences in how bias affects self-ratings of learning gains. As shown in Table 5, these SERU-S results give us an initial appreciation of the regularities and patterns in retrospective pretest data and perhaps some of the complexity its use will entail.

Examining campus differences

The results presented thus far represent the 12,500 seniors who entered as freshmen across the University of California, that is, without explicit reference to individual campuses. But

Table 4 Percent rating skills as "Very Good" or "Excellent" by one factor when controlling for other four factors (unweighted average of sixteen comparisons)

	Quant (%)	Oral (%)	Writing (%)	Reading (%)	Thinking (%)	Field (%)
UC GPA						
Began						
GPA < 3.2	25	17	21	21	20	6
GPA ≥ 3.2	33	17	25	22	27	6
Now						
GPA < 3.2	40	53	57	64	69	70
GPA ≥ 3.2	45	56	62	73	78	80
Gain						
GPA < 3.2	+15	+36	+36	+43	+49	+64
GPA ≥ 3.2	+12	+39	+37	+51	+52	+73
Difference in gains	−3	3	2	8	3	10
Field of study						
Began						
Science	34	16	25	21	26	6
Not science	24	18	22	21	21	6
Now						
Science	58	54	50	65	71	75
Not science	27	55	69	72	77	75
Gain						
Science	+24	+38	+26	+44	+45	+68
Not science	+3	+37	+47	+50	+56	+69
Diference in gains	−21	−1	22	6	10	0
Ethnicity						
Began						
Asian	29	15	22	19	22	6
Not Asian	28	19	25	23	25	6
Now						
Asian	40	48	52	59	65	69
Not Asian	45	62	67	77	82	81
Gain						
Asian	+11	+33	+31	+41	+44	+63
Not Asian	+16	+43	+42	+54	+57	+74
Difference in gains	5	10	11	13	13	12
Immigrant						
Began						
Immigrant	28	17	19	18	19	6
Not immigrant	30	17	27	24	28	6
Now						
Immigrant	42	57	57	67	72	73
Not immigrant	43	53	62	70	76	77
Gain						
Immigrant	+14	+40	+38	+49	+52	+66
Not immigrant	+13	+35	+35	+45	+49	+71

Table 4 continued

	Quant (%)	Oral (%)	Writing (%)	Reading (%)	Thinking (%)	Field (%)
Difference in gains	−1	−5	−3	−3	−3	5
Gender						
Began						
Male	31	15	22	20	26	7
Female	26	19	24	22	21	5
Now						
Male	49	54	59	68	77	76
Female	36	56	60	69	71	74
Gain						
Male	+17	+38	+37	+48	+51	+69
Female	+10	+37	+36	+47	+50	+69
Difference in gains	−8	−1	0	−1	−1	0

what about campus level data? Are there outliers that influence the aggregated data on self-reported learning gains?

The 2008 data set included survey responses from eight undergraduate campuses at the University of California. A ninth undergraduate campus, UC Merced, first enrolled students in 2006 and was therefore not included in our analysis. Among these eight campuses, Table 6 illustrates the differences and indicates why the display of such differences without further analysis is misleading. As can be seen in the top panel, two University of California campuses (campuses G and H) are clear outliers or "winners" with higher percentages of their seniors rating themselves as skillful or proficient than at other campuses.

However, simply adjusting for two broad differences in campus composition, Asian versus non-Asian and science versus non-science, eliminates entirely the apparent advantage of one of the campuses and substantially reduces it for the other. (Additional controls, e.g., for socioeconomic composition, would likely eliminate entirely the advantage in the second case.) Being able to demonstrate this provides a very practical application of our initial research findings on the social context of student self-ratings at the University of California. A more thorough examination of data at the campus level is beyond the scope of this initial study, but one we hope to pursue at a later date.

Conclusion

Compared to the Collegiate Learning Assessment, SERU-S offers a more nuanced approach in addressing the need for greater accountability for assessing and reporting learning outcomes in higher education, with significant considerations given to student and program complexity. There are solid connections between self-reports and students GPA and strong face validity of learning outcomes based on self-reports.

But the example of the apparent differences in learning outcomes across the undergraduate campuses of the University of California also illustrates the pitfalls and limitations of aggregated institutional level data, as well as the self-report data. Institutional level learning outcomes are confounded by much complexity unaccounted for with simple

Table 5 Percent seniors rating current skills as "Very Good" or "Excellent" by ethnicity, field of study, and UC GPA

	Asian		Not Asian	
	Science (%)	Not science (%)	Science (%)	Not science (%)
Specific field of study				
GPA < 3.2	62	63	76	78
GPA ≥ 3.2	76	75	85	83
Analytical and critical thinking				
GPA < 3.2	57	62	76	82
GPA ≥ 3.2	68	75	84	87
Reading				
GPA < 3.2	51	57	70	76
GPA ≥ 3.2	61	69	77	84
Writing				
GPA < 3.2	43	55	56	73
GPA ≥ 3.2	43	68	58	80
Oral presentation				
GPA < 3.2	45	46	60	62
GPA ≥ 3.2	50	50	63	63
Quantitative skills				
GPA < 3.2	47	25	60	27
GPA ≥ 3.2	60	29	65	26

controls at entry. Though tempting, we need to be careful in accepting at face value self-reports of learning and educational outcomes. SERU data shows signs of upward bias (social desirability, "halo" effect, etc.) inherent in self-reported data in institutional research (Gonyea 2005). The problem is further compounded by the fact that we already have evidence that the extent of bias is not uniform, as seen in the observed differences between Asian and non-Asian respondents.

On the other hand, these data, and the fact that they can be related to the extensive academic engagement data also collected on the SERU Survey as well as to the range of demographic and institutional data also available, offers an unprecedented opportunity to advance our understanding of the nature of self-reported learning outcomes in higher education and the extent to which these reports can to contribute as indirect but valid measures of positive educational outcomes at the research university.

Among our conclusions on the use of learning outcomes data:

1. While the SERU data are collected for entire campuses, the unique value of the census design is our ability to "drill down" to individual academic departments, student subpopulations, and other fine-grained "units of analysis." In examining patterns of learning outcomes, it will be particularly useful to do so at the level of student major (Chatman 2007) and to provide departments the ability to "triangulate" disciplinary-specific direct measures of learning with student self-reported gains. This can also pinpoint the areas that appear problematic, which allows targeted actions and interventions.

Table 6 Percent rating current skills as "Very Good" or "Excellent" by individual university of California campus before and after adjusting for differences in Asian versus non-Asian and science versus non-science composition

Campus	Unadjusted (%)	Adjusted (%)	Change (%)
Critical thinking			
A	72	71	−1
B	73	73	0
C	73	72	−1
D	73	71	−2
E	75	74	−1
F	77	77	0
G	83	74	−9
H	84	78	−6
Writing clearly			
A	60	59	−1
B	56	56	0
C	67	65	−2
D	59	58	−1
E	60	59	−1
F	61	59	−2
G	72	62	−10
H	72	61	−11
Field of study			
A	75	75	0
B	74	74	0
C	74	74	0
D	76	75	−1
E	74	75	1
F	75	75	0
G	83	75	−8
H	84	81	−3

2. Used properly, the extensive student self-reported indirect measures of learning outcomes should encourage greater attention to direct measures of student learning, but not serve as a substitute for such measures.
3. Conversely, "lowest common denominator" calculations of learning gains, such as deriving global outcome measures for an entire campus, especially without adjustment for student characteristics and compositional effects, will be less helpful, especially for encouraging campus self-improvement. In the VSA and elsewhere, it is precisely these kinds of global measures that are used, even though we know that such measures can be very misleading.

The census approach to collecting student self-reported learning gains offers the possibility of a different metric or unit of analysis, one that is predicated on institutional self-improvement. For example, of the 25 largest departments at a research university, how many have student ratings that meet a certain criterion? How many have demonstrated improvement in learning gains, as reported by students in their majors? The focus, in other words, would be at a level that is interpretable and more amenable to change.

All of these conclusions point to the complexity of any serious effort to understand the patterns of learning outcomes in major research universities. Tests such as the CLA are

blunt tools, creating questionable data that serves immediate political ends, but may offer skewed ideas on 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.

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 here regarding the CLA, the political momentum to generate a one-size-fits-all model is powerful. It already has 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 international version of PISA? The momentum toward creating an international "gold standard" learning outcomes test appears significant.

Anticipating pressure from lawmakers and stakeholder, some 331 universities that participate in the VSA are expected to report publicly their student's performance on one of three tests noted previously, with the CLA in a dominant market position. It may very well be that data and research offered in this and other studies on learning outcomes will be viewed as largely irrelevant in the push and pull for market position and political power. But universities who desire data for self-improvement may find that student surveys, if properly designed, offer a useful and cost-effective tool. They also may offer a means to combat simplistic rankings generated by CLA and similar tests.

References

Adelman, C. (2006). Border blind side. *Education Week*, *26*(11).

Arum, R., & Roska, J. (2008). *Learning to reason and communicate in college: Initial report of findings from the longitudinal CLA study*. New York, NY: Social Science Research Council.

Arum, R., & Roska, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago: University of Chicago Press.

Assessment of Higher Education Learning Outcomes. (2011). Project Update, Organisation for Economic Development and Cooperation. http://www.oecd.org/dataoecd/8/26/48088270.pdf.

Banta, T. (2006). Reliving the history of large-scale assessment in higher education. *Assessment Update*, *18*(4), 3–4, 15.

Banta, T. (2007). A warning on measuring learning outcomes. *Inside Higher Education*, January 26, found at: http://www.insidehighered.com/views/2007/01/26/banta.

Banta, T. (2009). *Assessment for improvement and accountability*. Provost's Forum on the Campus Learning Environment, University of Michigan, February 4, 2009.

Braun, H. (2008). *Viccissitudes of the validators*. 2008 Reidy Interactive Lectures Series, Portsmouth, NH. http://www.hciea.org/publications/RIL508_HB_092508.pdf. Last accessed January 24, 2012.

Chatman, S. (2007). *Institutional versus academic discipline measures of student experience: A matter of relative validity*. Berkeley: Center for Studies in Higher Education, University of California.

Consortium on Financing Higher Education (COFHE). (2008). *Assessment: A fundamental responsibility.* Found at: http://www.assessmentstatement.org/index_files/Page717.htm.

Gonyea, R. M. (2005). Self-reported data in institutional research: Review and recommendations. In P. D. Umbach (Ed.), *New directions for institutional research* (Vol. 127, pp. 73–89). San Francisco: Jossey-Bass.

Hill, L. G., & Betz, D. I. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation, 26*(4), 501–517.

Hosch, B. J. (2010). Time on test: Student motivation and performance on the college learning assessment: Implications for institutional accountability. Paper presented at the association of institution researchers conference, June 2, 2010. http://www.ccsu.edu/uploaded/departments/AdministrativeDepartments/Institutional_Research_and_Assessment/Research/20100601a.pdf.

Howard, G. S. (1980). Response-shift bias: A problem in evaluating interventions with pre/post self-reports. *Evaluation Review, 4*, 93–106.

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology, 64*, 144–150.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal invalidity in pretest-postest self-report evaluations and a re-evaluation of retrospective pre-tests. *Applied Psychological Measurement, 3*, 1–23.

Klein, S., Benjamin, R., & Shavelson, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review, 31*(5), 415–439.

Klein, S., Freedman, D., Shavelson, R., & Bolus, R. (2008). Assessing school effectiveness. *Evaluation Review, 32*(6), 511–525.

Klein, S., Kuh, G., Chun, M., Hamilton, L., & Shavelson, R. (2005). An approach to measuring cognitive outcomes across higher education Institutions. *Research in Higher Education, 46*(3), 251–276.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213–236.

Lam, T. C. M., & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *American Journal of Evaluation, 24*(1), 65–80.

Pike, G. R. (2006). Value-added measures and the collegiate learning assessment. *Assessment Update, 18*(4), 5–7.

Shulman, L. S. (2007). Counting and recounting: Assessment and the quest for accountability. *Change*, *39*(1), 28–35.

Spelling Commission on the Future of Higher Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. US Department of Education, September 26, 2006.

Taylor, P. T., Russ-Eft, D. F., & Taylor, H. (2009). Gilding the outcome by tarnishing the past. *American Journal of Evaluation, 30*(1), 31–43.