Research & Occasional Paper Series: CSHE.5.09

CSHE | Center for Studies in Higher Education

UNIVERSITY OF CALIFORNIA, BERKELEY

http://cshe.berkeley.edu/

SERU Project and Consortium Research Paper*

DECODING LEARNING GAINS
Measuring Outcomes and the Pivotal Role of the Major and Student Backgrounds

May 2009

Gregg Thomson and John Aubrey Douglass**

ABSTRACT

Throughout the world, interest in gauging learning outcomes at all levels of education has grown considerably over the past decade. In higher education, 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 viewed as a diagnostic tool for institutional self-improvement. This essay compares the methodology and potential uses of three tools for measuring learning outcomes: the Collegiate Learning Assessment (CLA), the National Survey of Student Engagement (NSSE), and the University of California's Undergraduate Experience Survey (UCUES). In addition, we examine UCUES 2008 responses of seniors who entered as freshmen on six of the 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. This initial analysis shows that campus-wide assessments of learning outcomes are generally not valid indicators of learning outcomes, and that self-reported gains at the level of the major are perhaps the best indicator we have, thus far, for assessing the value-added effects of a student's academic experience at a major research university. UCUES appears the better approach for assessing and reporting learning outcomes. This is because UCUES offers more extensive academic engagement data as well as a much wider range of demographic and institutional data, and therefore 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 contribute as indirect but valid measures of positive educational outcomes. At the same time, the apparent differences in learning outcomes across the undergraduate campuses of the University of California without controls for campus differences in composition illustrates some of the limitations of self-reported data.

In the US and throughout the world, interest in gauging learning outcomes at all levels of education has grown considerably 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 in some sense the quality and effectiveness, of colleges and universities. But perhaps most importantly, collecting and making public more and better *assessment* data about how and what students learn offer an important and relatively new means for institutional self-improvement.

* The SERU Project and Consortium is a collaborative of 15 major research universities based at the Center for Studies in Higher Education at UC Berkeley and including the administration of the SERU survey of undergraduates.

** Gregg Thomson is Executive Director of the Office of Student Research and Campus Surveys at UC Berkeley; John Aubrey Douglass is a Senior Research Fellow at the Center for Studies in Higher Education; both are co-PI's on the SERU Project and Consortium. Initial analysis and report preparation was conducted by Preeta Saxena of the UC Riverside Survey Research Center, under the direction of Steven Brint and David Crow, Associate Director of the UC Riverside Survey Research Center. Thanks to colleagues David Radwin, Steve Chatman, Cynthia Schrager, Elizabeth Berkes, and Dennis Hengstler for their insights regarding the use of SERU/UCUES data.

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 funding and 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 and extensive learning assessment program in U.S. higher education. 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).[1]

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": the Collegiate Assessment of Academic Proficiency (from ACT), the Measure of Academic Proficiency and Progress (from the Educational Testing Service), and the Collegiate Learning Assessment (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 Voluntary System of Accountability, 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."[2]

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."[3]

Regardless of the opposition to standardized testing to assess learning outcomes, the imperative to

measure and report on student learning outcomes for accreditation and public accountability remains strong. The University of California, for example, is implementing a comprehensive Accountability Framework, and it is expected that students' self-reported measures of learning will be included using data from the University of California Undergraduate Experience Survey – a survey developed by the UC community as part of the Student Experience in the Research University Project (henceforth referred to as SERU/UCUES).

But what is the best approach for assessing student learning outcomes?

This paper discusses three possible means for measuring learning outcomes for major research universities (including the University of California) and their strengths and weaknesses:

- The Collegiate Learning Assessment (CLA), which has emerged as the most visible of the standardized tests of student learning;

- The nationally prominent National Survey of Student Engagement (NSSE); and

- The SERU/UCUES survey currently used by the University of California and, more recently, at a number of other AAU public research universities.[4]

We also provide in this essay an exploration of SERU/UCUES self-reported learning gains. Previous research with SERU/UCUES data has documented both the striking demographic diversity of the undergraduate student body (Douglass, Roebken & Thomson, 2007; Douglass & Thomson, 2008) and the significant differences in student experience by field of study at the University of California (Chatman, 2007; Brint, Cantwell & Hanneman, 2008). Therefore, a major consideration in determining the best approach to measuring learning outcomes (more broadly, educational outcomes) at the University of California is that it should be both cost-effective and up to the task of addressing this demographic and disciplinary complexity.

Moreover, while some view the criteria for choosing how to measure learning outcomes in terms of the ability to assess the relative "value added" of institutions, our view is that decisions about the collection of student learning data should also be guided by the potential of this assessment to encourage institutional self-improvement.

A. CLA — The Negative Value of "Value Added"

Although the University of California initially rejected the use of standardized tests to assess learning outcomes, the increasing prominence and use of the Collegiate Learning Assessment by other colleges and universities (including, for example, the University of Texas system) suggests that the CLA is certainly worth a close look.

Developed by the Council for Aid for Education, the Collegiate Learning Assessment (CLA) offers a carefully developed 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. Test results derived from these samples provide an institution-wide measure of the institution's contribution or value-added to the development of its students' cognitive competencies or learning. Institutions can then be compared on the basis of their relative value-added performance.

The value of the CLA derives from two well-articulated 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 (Klein et al., 2005; Klein, Benjamin & Shavelson, 2007).

- Second, by using SAT scores as the control for initial student characteristics, given how well the CLA tests have been designed and validated as measures of general cognitive skills, it is possible on the basis of surprisingly small samples to calculate the difference between freshman and senior test performance and compare that difference to that predicted or expected on the basis of 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, Benjamin & Shavelson, 2007; Klein et al., 2008).

. . . 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 will simply be in the expected range.

Banta (2006, 2007, 2009) as well as other prominent higher education researchers (e.g., Pike, 2006) have questioned the CLA enterprise 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 will simply be in the expected range. The results are also sample-dependent in ways not recognized by CLA (for example, student motivation). Finally, the 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 general cognitive skills that are relevant to all students (Pike, 2006), but recent findings (Arum & Roska, 2008) suggest that CLA results are, to some extent, discipline-specific. Because of the cost and difficulty of evaluating individual student essays, the design of the CLA requires a rather small sample size (often 250 to 300 students) and thereby generates generalities about overall institutional effectiveness. There is very little if any useful information at the level of the major. The CLA might generate meaningful data in a small liberal arts college, but it appears of very limited use in large and complex 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).

The CLA 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, Benjamin & Shavelson, 2007; Klein et al., 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, this does not mean that it is appropriate for the University of California to commit its resources to this enterprise.

There are at least three reasons for rejecting the implementation of the CLA for institutional “accountability” at the University of California regardless of what (or who) one believes regarding the arguments for its validity.

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 this is a preposterous proposition for a large, complex research university like the University of California.

Exactly how would the statistically derived result (on the basis of a few hundred freshman and senior test-takers) or news 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? In reality, this news would surely generate “more heat than light” and could offer no guidance whatsoever in terms of institutional self-improvement.

Second, any approach to the assessment of student learning at the University of California that provides no ability to examine how well the university is doing in regard to its student populations from various backgrounds and life circumstances is incompatible with its core value of diversity and access.

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 validity, namely, what Braun identifies as "systemic validity":

> 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. (2008)

"Valid" or not, the successful promotion of a narrow standardized test "value-added" program of assessment in higher education promises little in the way of "useful information and constructive responses" while threatening "undue deterioration" elsewhere. Such a ranking system 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, then, we surmise that the net value of CLA's value-added scheme would be more negative than positive.

"Valid" or not, the successful promotion of a narrow standardized test "value-added" program of assessment in higher education promises little in the way of "useful information and constructive responses" while threatening "undue deterioration" elsewhere.

B. Student Self-Reports of Learning Gains on NSSE

Over the last decade or so, the National Survey of Student Engagement (NSSE) has grown remarkably in its use among a great variety of higher education institutions, although most predominately in liberal arts colleges. The most recent annual report notes:

> Like the speaker who "needs no introduction," NSSE may well have achieved an eminence that needs no foreword. The acronym is everywhere: on institutional Web sites and the lips of parents and students selecting a college; the pages of *USA TODAY*, the *Chronicle of Higher Education*, *Change* magazine, and the *New York Times;* the 2006 report from the National Commission on the Future of Higher Education, and now on the template for the Voluntary System of Accountability. … In fact, go to Google and you'll find "about 299,000" entries that deal with NSSE. (National Survey of Student Engagement, 2008b)

Established a decade ago and promoted as a constructive alternative to invidious college ranking schemes (especially the *US News and World Report* rankings), NSSE is *the* obvious source for useful survey-based information on undergraduate learning outcomes. Or so it would seem.

Thirty years ago Pace (1979, 1984) initiated the systematic collection of undergraduate student experience data with the College Student Experience Questionnaire (CSEQ), and it included items on self-reported educational outcomes. In 1999 Pace's format for these items was incorporated verbatim into the National Survey of Student Engagement (NSSE) and ten years later remains the basis for a 16-item section on NSSE. The question reads, "To what extent has your experience at this institution contributed to your knowledge, skills, and personal development in the following areas?" The possible answers are "Very much," "Quite a bit," "Some," and "Very little."

Asking about educational outcomes in the way NSSE does, that is, without reference to either a beginning point or other standard and with vague response categories, is fundamentally flawed. Responses are subject to very significant "halo" effects (Pike, 1999; Wells, 1907) and are not valid indicators of actual learning or educational gains (Gonyea, 2005; Pascarella, 2001).

Researchers using variations of the NSSE approach have failed to find any valid relationship between student self-reports of learning gains measured this way and actual gains. Pike (1993) used only three response options and, more recently, Bowman (2009), whose study using four response options, found no correlation between self-reported gains and independently measured growth in the freshman year. (Interestingly, the study by Anaya (1999) that did find that self-reports were a valid measure of learning

used an explicit five-point scale of change in skills: (1) much weaker, (2) weaker, (3) no change, (4) stronger, and (5) much stronger.)

What is stunning, however, is that a full decade of NSSE activity has produced no research that establishes the validity of the NSSE approach for measuring gains in learning or other educational outcomes. Two early local studies (Belcheir, 2001, 2003) produced confusing and largely uninterpretable results, and a more recent larger and methodologically sophisticated study (Carini, Kuh & Klein, 2006) found no relationships between senior self-reports of educational gains and various independent measures.

In addition to the lack of demonstrated validity, the NSSE educational outcomes results are problematic on even the intuitive or descriptive level. The NSSE freshman-senior sample design should, one would expect, showcase significant learning gains by comparing the freshman and senior results. However, this is not the case at all. For freshmen on the 2008 NSSE the average across the 16 educational outcomes items reporting gains of "Quite a bit" or "Very much" is 63%; for seniors the figure is 65% (National Survey of Student Engagement, 2008b).

In addition to the lack of demonstrated validity, the NSSE educational outcomes results are problematic on even the intuitive or descriptive level.

In other words, using the NSSE-style items, it appears superficially that there are significant educational gains made in the freshman year and then almost no additional gains over the next three years! Not exactly a case for positive learning outcomes or institutional "value-added." Finally, there is no mention of the 16-item educational outcomes section of NSSE or any of the items from it in the most recent comprehensive 50-page report on the 2008 NSSE results (National Survey of Student Engagement, 2008b).

Given the prominence of the NSSE enterprise, its decade-long adherence to a fundamentally flawed approach constitutes a glaring gap (literally, in its annual reports) and missed opportunity for our understanding of the nature of learning and other educational outcomes in higher education. Because of its freshman-senior sample design (as opposed to the SERU/UCUES census approach), NSSE would probably not provide the scope of learning outcomes data needed by the University of California even with the most valid survey items. But lacking valid measures of learning outcomes entirely, NSSE, thus far designed, clearly "fails the test."

C. Why SERU/UCUES May Succeed

The major research universities that are part of the Voluntary System of Accountability have identified SERU/UCUES as one of four nationally recognized surveys for institutional accountability.[5] The overwhelming value of SERU/UCUES for the comprehensive informational needs of the University of California – and by extension other large-scale research universities – is its census (plus module) design that provides data down to the level of individual academic program and student subpopulations of interest.

In terms of assessing learning outcomes specifically, however, SERU/UCUES also offers an innovative approach that sets it apart from conventional undergraduate surveys. This approach is drawn from the field of program evaluation where 30 years ago the work of Howard (1980; Howard et al., 1979; Howard & Dailey, 1979) and then others challenged the conventional wisdom that the most valid way to measure program effects or gains is to use a pretest-posttest design. Because of what Howard identified as "response-shift bias", program participants are likely to have a more informed frame of reference as a consequence of their experience in the program, often making posttest evaluations of their proficiencies or knowledge both lower and more accurate than their pretest evaluations.

With this insight, assessment of program or treatment effects did not need to rely as heavily on the more costly pretest-posttest evaluation design and could often substitute the retrospective posttest design. More recent research, however, has concluded that earlier views of the magnitude of response-shift bias were exaggerated (Wilson & Lipsey, 2001), and that the retrospective posttest design may produce ratings that are more biased than prospective ratings (Hill & Betz, 2005; Taylor, Russ-Eft, & Taylor, 2009). Not surprisingly, inflating the difference between retrospective and current self-ratings is associated with social desirability (Hill & Betz, 2005). Retrospective pretest ratings may be lowered because of motivational or systematic cognitive bias such as self-enhancement, implicit theory of change, and effort justification (Taylor, Russ-Eft, & Taylor, 2009).

On the other hand, when comparing the retrospective pretest method with the perceived change method (similar to the NSSE approach) and the post plus perceived change method with teachers reporting change in instructional practices, Lam & Bengo (2003) found that the retrospective pretest method produced the least "satisficing" (Krosnick, 1991) and responses on the basis of social desirability, while the perceived change method produced the most.

Several observations and generalizations emerge from the practice of the retrospective pretest method in program evaluation and other fields. The method is especially useful if capturing accurately how change is experienced subjectively by program participants is relevant. 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. Finally, if the costs for overestimating program effects are not great, the advantages of using this approach can offset the potential biases.

All of these conditions would seem to apply to any large-scale effort to assess and report learning outcomes in higher education. A method that allows us to capture accurately how different populations of students (e.g., students in different majors) characterize their own learning gains at the University of California and under what conditions should contribute considerably to our potential understanding of the complexities of learning outcomes.

And 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.

. . . a large-scale census design allows us to amass tremendous amounts of learning and other educational outcomes data at a fraction of the cost of any other method, thereby providing an opportunity to conduct analyses to determine how self-reports are affected by "inflationary bias" and the extent (or not) that they are useful in validating and reporting learning outcomes.

Granted, student self-reports of learning are only indirect indicators and are clearly favorably biased ones at that. On the other hand, the large-scale census design allows us to amass tremendous amounts of learning and other self-reported educational outcomes data at a fraction of the cost of any other method, thereby providing an opportunity to conduct analyses to determine how self-reports are affected by "inflationary bias" and the extent (or not) that they are useful in validating and reporting learning outcomes.

Therefore, in 2004 the original SERU/UCUES survey instrument included the retrospective pretest or "then" and "now" self-report methodology to measure educational outcomes for University of California undergraduates rather than adopting the rating of improvement approach used by NSSE and CSEQ. Initially, for 14 educational outcomes, students were asked to assess their skills and proficiencies on a seven-point scale (Very Poor, Poor, Fair, Good, Very Good, Excellent, Expert), both when they started at the University of California and currently.

More recent versions of SERU/UCUES use a six-point scale (Very Poor, Poor, Fair, Good, Very Good, Excellent), and the latest instrument has ratings of 21 different educational outcomes (and students assigned to the Student Development module rate an additional six outcomes).

Given the SERU/UCUES census design, therefore, we now have an incredibly 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 gives us the ability to help assess and control for the tendency to exhibit improvement biases.

An earlier study (Thomson, 2006) examined the SERU/UCUES educational gains data for University of California, Berkeley freshmen, sophomores, juniors, and seniors and found that the student self-reports demonstrated clear evidence of response-shift bias and/or self-enhancement bias: With each year in school, ratings of "when you started' were lower.

But the results also showed a high degree of stability and "reasonableness" in that reported gains were modest and selective (i.e., respondents did not report gains in all areas). Most importantly, the magnitude of reported gains for juniors and seniors differed by domain and field of study, suggesting that the student self-reports, though biased upward, did appear to reflect different patterns of learning.

D. The Current Study

The research reported here examines the responses of seniors who entered as freshmen on six of the educational outcomes self-reports on SERU/UCUES 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. Omitting respondents with missing data on UC GPA, gender, race/ethnicity, immigrant generation, or major, the study has about 12,500 sets of responses.

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 1 shows how University of California seniors assess their learning gains in each of the six areas. While seniors are more likely to rate themselves as 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.

If the SERU/UCUES senior self-reports of learning gains have validity, then we should observe a relationship with another assumed measure of learning, namely college GPA. The relationship between student self-reports and overall cumulative UC GPA is examined in Table 2.

Table 2. Percent Rating Skills as "Very Good" or "Excellent" Across Six Domains by Current Cumulative UC GPA Category

Quantitative	Began	Now	Gain
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	**Began**	**Now**	**Gain**
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	**Began**	**Now**	**Gain**
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	**Began**	**Now**	**Gain**
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 Thinking	**Began**	**Now**	**Gain**
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	**Began**	**Now**	**Gain**
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%

This way in which the relationships of UC GPA and student self-reports vary by skill area does suggest that student assessments using the SERU/UCUES approach have some degree of validity as indicators of learning outcomes. Specifically, the relationship is weakest for quantitative skills (gains are actually the lowest for the highest GPA students) and oral presentation skills, skill areas less uniformly related to academic achievement across all majors.

Conversely, it is strongest for critical and analytical thinking and field of study. As shown in Table 3, senior self-reports are also related significantly to student demographics and field of study.

Our first look at the results indicates 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 next examined the effects of the factors in combination. To do this, student responses were analyzed using a 2 X 2 X 2 X 2 X 2 design:

- UC GPA < 3.2 versus UC GPA >= 3.2
- Science (STEM disciplines) versus non-science
- Immigrant (both parents not born in US) versus non-immigrant
- Male versus female
- 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.

The 2 X 2 X 2 X 2 X 2 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*.

Gains, after controlling for other factors, 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.

Table 3. Senior Self-Reports by Ethnicity, Demographics and Field of Study

ETHNICITY		Quant	Oral	Writing	Reading	Thinking
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		Quant	Oral	Writing	Reading	Thinking
Began	Student Not Born in U S	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 U S	41%	42%	43%	54%	56%
	Parent(s) Not Born in U S	36%	44%	51%	58%	59%
	Both Parents Born in US	36%	50%	66%	72%	78%
Gains	Student Not Born in U S	+13	+24	+26	+34	+36
	Parent(s) Not Born in U S	+11	+22	+29	+34	+37
	Both Parents Born in US	+10	+23	+32	+38	+42

FIELD OF STUDY		Quant	Oral	Writing	Reading	Thinking
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
	Social Sciences	+06	+31	+40	+44	+42
	Humanities	-03	+28	+42	+44	+42

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.

Our approach here, of course, underestimates the full impact of various factors on senior 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/UCES 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.

E. Campus Differences and Accountability

The results presented here represent the 12,500 seniors who entered as freshmen across the University of California, that is, without reference to individual campuses. There is perhaps a natural curiosity to compare the senior self-reports of educational outcomes across campuses, and for public accountability, perhaps even an imperative to do so.

Table 4. Percent Rating Skills as "Very Good" or "Excellent" by One Factor When Controlling for Other Four Factors (Unweighted Average of Sixteen Comparisons)

UC GPA		Quant	Oral	Writing	Reading	Thinking	Field
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		Quant	Oral	Writing	Reading	Thinking	Field
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%
Difference in Gains		-21%	-1%	22%	6%	10%	0%

ETHNICITY		Quant	Oral	Writing	Reading	Thinking	Field
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		Quant	Oral	Writing	Reading	Thinking	Field
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%
Difference in Gains		-1%	-5%	-3%	-3%	-3%	5%

GENDER		Quant	Oral	Writing	Reading	Thinking	Field
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%

Table 5. Percent Seniors Rating Current Skills as "Very Good" or "Excellent" by Ethnicity, Field of Study, and UC GPA

SPECIFIC FIELD OF STUDY

	Asian		Not Asian	
	Science	Not Science	Science	Not Science
GPA < 3.2	62%	63%	76%	78%
GPA >= 3.2	76%	75%	85%	83%

ANALYTICAL AND CRITICAL THINKING

	Asian		Not Asian	
	Science	Not Science	Science	Not Science
GPA < 3.2	57%	62%	76%	82%
GPA >= 3.2	68%	75%	84%	87%

READING

	Asian		Not Asian	
	Science	Not Science	Science	Not Science
GPA < 3.2	51%	57%	70%	76%
GPA >= 3.2	61%	69%	77%	84%

WRITING

	Asian		Not Asian	
	Science	Not Science	Science	Not Science
GPA < 3.2	43%	55%	56%	73%
GPA >= 3.2	43%	68%	58%	80%

ORAL PRESENTATION

	Asian		Not Asian	
	Science	Not Science	Science	Not Science
GPA < 3.2	45%	46%	60%	62%
GPA >= 3.2	50%	50%	63%	63%

QUANTITATIVE SKILLS

	Asian		Not Asian	
	Science	Not Science	Science	Not Science
GPA < 3.2	47%	25%	60%	27%
GPA >= 3.2	60%	29%	65%	26%

Table 6 illustrates what University of California campus differences look like and why the display of such differences without further analysis is misleading. As can be seen in the top panel, two University of California campuses 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 vs. non-Asian and science vs. 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.

F. Conclusion

Compared to the Collegiate Learning Assessment (CLA) and the National Survey of Student Engagement (NSSE), SERU/UCUES appears the better approach in addressing the need for greater accountability for assessing and reporting learning outcomes in higher education. But the example of the apparent differences in learning outcomes across the undergraduate campuses of the University of California illustrates the obvious pitfalls and limitations of the self-report data.

Though tempting, we cannot accept self-reports of learning and educational outcomes at face value. The UCUES/SERU data have all the problems of upward bias (social desirability, "halo" effect, etc.) inherent in self-reported data in institutional research (Gonyea, 2005).

The problem is compounded by the fact that we already have evidence that the extent of bias is not uniform, i.e., 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/UCUES 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.

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.

CRITICAL THINKING

Campus	Unadjusted	Adjusted	Change
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

Campus	Unadjusted	Adjusted	Change
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

Campus	Unadjusted	Adjusted	Change
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%

Our efforts here should be informed by the following:

(1) While the UCUE/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 the cost-effective externally generated SERU/UCUES survey data.

(2) Used properly, the extensive SERU/UCUES student self-reported indirect measures of learning outcomes should encourage greater attention to direct measures of student learning, not serve as a substitute for such measures. SERU/UCUES demonstrates that extensive individual student data can be collected electronically relatively inexpensively no matter how large the university. Large-scale use of electronic portfolios may be more feasible than generally thought (Banta, 2009).

The time has come for institutional researchers and analysts at the University of California to take full advantage of the tremendous amount of retrospective pretest data on educational outcomes that we have available from SERU/UCUES.

(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 Voluntary System of Accountability (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.

Conventionally, of course, with the use of sample surveys such as NSSE, only institution-wide statistics are available. SERU/UCUES 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 their majors? The focus, in other words, would be at a level that is interpretable and more amenable to change.

Our conclusion: The time has come for institutional researchers and analysts at the University of California to take full advantage of the tremendous amount of retrospective pretest data on educational outcomes that we have available from SERU/UCUES. We should extend our inquiry to the full array of 21 educational outcome items in the core, examine the data across the full range of undergraduate cohorts and subpopulations of interest, and identify and encourage any number of more focused "validity" studies.

We are optimistic that such efforts will significantly advance our understanding of educational outcomes and help facilitate the improvement of teaching and learning at the research university. To this we should be held accountable.

REFERENCES

Adelman, C. (2006). Border blind side. *Education Week*, 26 (11), November 8.

Anaya, G. (1999). College impact on student learning: Comparing the use of self-reported gains, standardized test scores, and college grades. *Research in Higher Education*, 40, 499-526.

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

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.

Belcheir, M. J. (2001). What predicts perceived gains in learning and in satisfaction? Report No. BSU-RR-2001-02). Boise, ID: Office of Institutional Advancement (ERIC Document Reproduction Service No. ED480921).

Belcheir, M. J. (2003). Student academic and personal growth while at Boise State: A summary of 2002 National Survey of Student Engagement findings. (Report No. BSU-RR-2003-03). Boise, ID: Office of Institutional Advancement (ERIC Document Reproduction Service No. Number ED480934).

Bowman, N. A. (2009). *Can first-year college students provide accurate self-reports about their learning and development?* Unpublished manuscript, University of Notre Dame.

Braun, H. (2008). *Viccissitudes of the Validators*. 2008 Reidy Interactive Lectures Series, Portsmouth, NH.

Brint, S, Cantwell, A. M., & Hanneman, R. (2008). The two cultures of undergraduate academic engagement. *Research in Higher Education*, 49(5), 383-402.

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

Chatman, S. (2007). *Institutional Versus Academic Discipline Measures of Student Experience: A Matter of Relative Validity*. Center for Studies in Higher Education, University of California, Berkeley.

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

Douglass, J.A., Roebken, H. & Thomson, G. (2007). *The immigrant university: Assessing the dynamics of race, major and socioeconomic characteristics at the University of California.* Center for Studies in Higher Education, University of California, Berkeley

Douglass, J.A. and Thomson, G. (2008). *The poor and the rich: A look at economic stratification and academic performance among undergraduate students in the United States.* Center for Studies in Higher Education, University of California, Berkeley.

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

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

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 pretests. *Applied Psychological Measurement*, 3, 1-23.

Klein, S., Benjamin, R., and 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.

National Survey of Student Engagement (2008a). *Frequency Distributions: 2008.* Indiana University Center for Postsecondary Research. Bloomington, IN

National Survey of Student Engagement (2008b). *Promoting engagement for all students: The imperative to look within. 2008 Results*. Indiana University Center for Postsesondary Research. Bloomington, IN

Pace, C. R. (1979). *Measuring the Outcomes of College*. San Francisco: Jossey-Bass
.

Pace, C. R. (1984). *Measuring the quality of college student experience: An account of the development and use of the College Student Experiences Questionnaire*. Los Angeles: Higher Education Research Institute.

Pascarella, E. T. (2001). Using student self-reported gains to estimate college impact: A cautionary tale. *Journal of College Student Development*, 42, 488-492.

Pike, G. R. (1993). The relationship between perceived learning and satisfaction in college: An alternative view. *Research in Higher Education*, 34, 23-40.

Pike, G. R. (1999). The constant error of the halo in educational outcomes research. *Research in Higher Education* 40, 61-86.

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*. January-February.

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.

Thomson, G. (2006). New developments in the assessment of student development and proficiences. Paper presented at the annual meeting of ACPA, Indianapolis, IN.

Wells, F. (1907). A statistical study of literary merit. *Archives of Psychology*, 7, 5-30.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods*, 6, 413-429.

NOTES

[1] The Spellings Commission was announced 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).

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

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

[4] These include Florida, Michigan, Minnesota, Pittsburgh, Rutgers, and Oregon.

[5] 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).