

BenjaminCAE

Assessment Versus Accountability in Higher Education: Notes for Reconciliation<sup>1</sup>

Roger Benjamin and Stephen Klein

Draft

Not To Be Quoted

March 2006

## Introduction

There is confusion about what kind of assessment is appropriate in higher education. There is also misunderstanding over the relationship between assessment and accountability. Even when institutions and state policymakers use similar assessment information they do so for different purposes. Faculty are focused on improving educational programs while state leaders are interested in holding their public higher education institutions accountable for their performance.

Our goal is to improve our understanding of assessment and accountability and the relationship between them. We do so through the following steps: First, we present a rationale for a set of assessment principles to implement in higher education. Second, we argue that states have an important role in assuring accountability of their institutions of higher learning but that they must avoid a number of land mines if they wish to engage their institutions successfully. Moreover, to do so, basic rules in comparative methodology must be adhered to in the use of institutional level data for any aggregated comparisons. Third, we present an example of the application of these assessment principles in a set of testing activities we find useful elements for any strategy of rapprochement between higher education

---

<sup>1</sup> The authors thank Catherine Augustine, RAND, and Marc Chun, Council for Aid to Education for comments on an earlier draft.

institutions and state-based authorities. We argue that assessment measures based on the principles stand the best chance to fill the gap between the two groups. These measures should be organized with other measures to form an assessment indicator system. Fourth, the terms of engagement between institutions and state-level policymakers must be carefully worked out.

### The Focus: Public or Private Institutions

The need to improve our understanding of assessment and accountability cuts across public and private colleges and universities. Regional accreditation groups now require evidence of student learning success for the public and private institutions they accredit. Private and public colleges are frequent targets of Congressional concerns about the cost of undergraduate education. Such concerns occasionally translate into proposed legislation to deal with the problem. Most recently, the Commission on the Future of Higher Education, established by Secretary of Education Spellings, has brought accountability issues to public attention. Because of concerns about costs and quality, private as well as public institutions need to demonstrate that they add value, that they produce successful educational outcomes. Boards of trustees of private and public institutions now increasingly call for evidence of success. Finally, both public and private institutions are interested in using assessment to improve teaching and learning. Many of the points raised here are therefore relevant to both public and private institutions. However, we feature public institutions in this study because they are under the direct authority of governors, state legislatures, and state-based commissions of higher education directly charged to hold public institutions accountable.

### The Argument

The “public” (taxpayers, legislators, governors) wants to be assured that their college students are receiving a quality education. This interest in accountability is fueled by the same factors that have led to higher tuitions, namely shrinking state budgets and the increasing cost of higher education.<sup>2</sup> In the past, institutions relied on accreditation reviews and various types of actuarial data, such as graduation and minority access rates, to demonstrate quality. That approach is no longer adequate for colleges just as it no longer sufficient for K-12 education (as evidenced by No Child Left Behind legislation and the emphasis on statewide testing of students).

---

<sup>2</sup> Jones (2003).

The public wants to know whether its education institutions are helping students acquire the knowledge, skills, and abilities they will need when they graduate.<sup>3</sup> In addition, policy makers increasingly want to know how much students actually have learned, not how much they believe that they have learned. Forty-four states have established accountability systems<sup>4</sup> for higher education. Within this group 27 states feature “report cards” that attempt to benchmark student learning outcomes.<sup>5</sup> Thus, the measure of quality has been expanded beyond accreditation and actuarial data to include evidence that learning goals have been met. Seat time, course grades, and graduation rates are no longer sufficient. In short, the public is increasingly asking its colleges and universities to show that acceptable progress has been made in student learning.

To satisfy this demand for accountability, higher education institutions need to demonstrate that their students have acquired important skills and knowledge in addition to achieving other goals such as graduating, achieving necessary prerequisites for professional schools, and gaining employment. Institutional ratings, student and faculty surveys, and other indirect proxies are just not sufficient. The only credible way to show such learning is to test them over what they are supposed to know and be able to do. Instead, direct measures of outcomes are needed. Colleges and universities already assess students, but hardly ever for the purpose of demonstrating the value the institution adds to a student’s knowledge and skills. At least until recently, their reasons for testing have had nothing to do with accountability. Instead, they test incoming students to make sure they have the skills that are needed to do their course work. Those who do not have sufficient skills are generally placed in remedial programs. In addition, some colleges administer tests at the end of the sophomore year to make sure students are ready for their upper division studies. These so-called “rising junior” exams, like the initial placement tests, focus on basic reading, writing, and math skills. These tests are focused on the individual student without attempting to measure the contribution of the institution to student learning.

---

<sup>3</sup> Immerwahl (2000).

<sup>4</sup> Burke and Minassians (2002).

<sup>5</sup> Naughton, et al., (2003). Naughton, et al. recorded 227 indicators related, in some way, to student learning. See also *Measuring Up* (2002) which gives student learning an incomplete for all the states. The regional accrediting association now require evidence of the quality of student learning. See the web page based discussion this topic of the Western Association of Schools and Colleges (WASC). See also the publications and activities devoted to this subject by the national accrediting body the Council for Higher Education Assessment (CHEA).

Some colleges are now expanding their testing programs to include assessing other abilities, such as critical thinking skills, that are central to the college's mission but cut across academic majors. College administrators see this as a way to demonstrate the beneficial effects of the educational experiences at their institutions to prospective students and their parents.<sup>6</sup> Nevertheless, most institutions continue to rely on their faculty to assess their students' content knowledge and skills. This is fine with the faculty who generally believe they already provide sufficient and appropriate assessments of student learning. They use midterm and final exams, term papers, classroom participation, and other evidence to assign grades. And, they feel these grades reflect how much students learn in their courses.<sup>7</sup> Unfortunately, grades of professors are idiosyncratic. Two courses with the same title may cover different content areas, even at the same college. There also are large differences in grading standards among professors across colleges. B-work at one school may correspond to A- or C-work at another institution. The same is true across professors within an institution.<sup>8</sup> There also has been substantial grade inflation over time.<sup>9</sup> Hence, professor assigned grades cannot be relied on to provide a valid measure of whether the students in one graduating class are more or less proficient than those in another class or at another college. Nor are value added comparisons of the contributions of institutions to growth in student learning made. Some other metric is needed.

The search for another index has led some colleges to experiment with portfolios, grades in capstone courses, or other institution-specific indicators of learning (Banta, 1996). However, all of these measures have the same fundamental limitation as regular course grades, namely, the absence of a way to reliably and validly interpret scores outside of the context of a particular course or school at a given point in time. To correct that problem, the measures have to be administered under the same standardized conditions to everyone and the scores obtained have to be adjusted for possible variation in average question difficulty, reader leniency, and other factors. Locally constructed measures, like professor course grades or

---

<sup>6</sup> The number of colleges developing general education programs is rising led by national associations of higher education such as the Association of American Colleges and Universities (AAC&U). See *General Expectations: A New Vision for Learning* (2002). And most colleges embed the goal of increasing these skills in their mission statement. See also Immerwahr (2000) for discussion of the expectations of parents and students about higher education.

<sup>7</sup> Moreover, although attendance at meetings on higher education assessment held by groups such as the Association of American Colleges & Universities (AAC&U) have increased in recent years, we believe most faculty remain skeptical of the kind of assessment discussed here.

<sup>8</sup> For example, (Klein, et. al, 2005) developed a method to deal with the problem of widely divergent grading patterns across institutions. They converted GPAs within a school to z-scores and then used a regression model (that included the mean SAT score at the student's college) to adjust the correlations with GPAs for possible differences in grading standards among colleges.

<sup>9</sup> For example, Harvard, where, until this past year, 90 percent of the undergraduate students, were deemed honors students, is, like other institutions, attempting to cope with the problem of grade inflation. ( Dondio, 2005).

portfolios, do not have these essential features and therefore cannot be used for making valid comparisons within institutions over time or for comparisons among institutions at a single point in time.

Those limitations are not present with the measures that are used for statewide K-12 testing (such as the Stanford-9, Iowa Tests of Basic Skills, and the National Assessment of Educational Progress (NAEP)), college and graduate school admission decisions (such as the SATs, ACTs, GREs, and LSATs), or licensing exams in the professions (such as accountancy, law, medicine, and teaching). Thus, when results really matter, such as for high-stakes decisions about individuals, procedures are used that help to eliminate the effects of extraneous factors, such as who drafted the questions or scored the answers.

### The Role of the State

States have a legitimate and critical role in assuring accountability in their higher education institutions.

Many states set objectives for:

- educational proficiency levels to be achieved by entering students
- participation rates by ethnic/racial groups
- minimum passing scores for law, medicine and other professional schools
- numbers of graduates in particular fields to be achieved such as teaching, nursing, and technology fields.<sup>10</sup>

The states also provide the instructional budgets for public undergraduate education and infrastructure support, including buildings, library, and scientific equipment. States clearly have a right and a responsibility to require accountability from the institutions they support. Why, then, are we not further ahead in developing assessment systems of student learning that work from the point of view of the institutions and the states?

The problem is that the assessment measures used at the institutional level differ from most of the accountability measures states focus on. First, the concept of accountability must be specified. Most broadly, in the context of higher education, accountability can be defined as the extent to which public

---

<sup>10</sup> See the master plans of state higher education coordinating commissions or governing boards of state higher education systems, for example, of Texas or Nevada.

higher education institutions meet the goals set for them by the state. (In the best case these goals are mutually agreed to by both parties.) Just as faculty and institutions set assessment goals for a variety of purposes, states set accountability goals for different purposes. Most states desire accountability for prudent use of resources, or at the very least, absence of fraud. Some state leaders demand evidence of increased participation, retention and graduation rates for underrepresented groups. Still others, an increasing number, want to be assured that students have gained knowledge and skills from their educational experiences. Approaches and measures of student learning favored by faculty differ from those used by state leaders. Because of the growing interest in student outcomes, we are focused on this last goal of state-based accountability, evidence of student learning outcomes.<sup>11</sup>

Approaches to student learning outcomes by faculty have the following characteristics: Their goals are to improve curriculum and pedagogy as well as set targets for students:

- They focus on individual students, departments, or institutions but are not focused on inter-institutional comparisons;
- Are content rich, tailored to the context of the institution and generated by faculty themselves and are often time intensive and costly;
- Because the emphasis is on content, they tend not to be replicable from one institution to the next.

In comparison, state-based approaches are:

- focused on accountability objectives;
- aggregated at the regional or state level and ideally replicable and comparable across institutions;
- focused on indirect proxies of student learning outcomes such as the percentage of passing rates for teaching examinations,
- nurses, and other professional school examinations; number and percentage of students that take the graduate record examination (GRE); retention and graduation rates;
- not rich in content or tailored to the context of the individual institutions and not developed by faculty;
- cost effective, making use of existing data .

The result is a disconnect between the faculty/institutions on the one hand and the state on the other.

---

<sup>11</sup> Naughton, et al. (2003).

## Comparative Methodology Obstacles

This disconnect is made worse because a number of state-based comparisons violate comparative methodological principles. The attempt at comparing states at the K-12 level<sup>12</sup> is now being extended to the higher education level.<sup>13</sup> It is not easy to make direct comparisons among states on student learning outcomes. Such comparisons are fraught with methodological hurdles, some of the more important of which are listed below:<sup>14</sup>

- States differ dramatically in demographic and social-economic characteristics so that direct comparisons, say, between California and Rhode Island about the mean proficiency levels attained in math or reading make little sense.<sup>15</sup>
- Comparing aggregated scores at the state level, rather than higher education institutions, makes little sense because such comparisons assume there are no differences in effects across individual higher education institutions within a state. If a state's scores go up (or down), is it due to one or all institutions? If a state's scores stay the same, is it because all the institutions functioned the same way or did the effects at one institution offset those at another? This is the fundamental flaw that statisticians call "aggregation bias" or the ecological fallacy problem.<sup>16</sup>
- The use of a variety of indirect, proxy measures is problematic for several reasons. Comparing GRE scores across states makes no sense because of concerns about selection bias, e.g., a state may have a large or small number of students who take the GRE exam; the context that drives the number differs substantially from state to state. Use of passage rates on licensing exams, such as

---

<sup>12</sup> See the National Assessment of Educational Progress (NAEP) carried out by the National Center for Educational Statistics, United States, Department of Education.

<sup>13</sup> Measuring Up (2002; 2004).

<sup>14</sup> The literature on comparative methodology in social inquiry is instructive. See Przeworski & Teune (1970) for the logic of selecting most similar or most different designs; see Ragin (1989) for the logic of comparable case selection and analysis; cf Collier (1991). We note also that many state-based leaders of higher education departments or coordinating boards are well aware of these methodological hurdles. They are faced with the practical necessity of making comparative measures, often as a function of legislative mandate. They thus are forced to use data and measures created for many different purposes.

<sup>15</sup> International comparisons must be cautioned for the same reasons. To compare the national unit the United States with individual European nations throws apples and oranges together. Indeed the international comparisons of the TIMSS (National Center for Education Statistics) are problematic because it compares the United State with, for example, Finland and Norway in terms of levels of achievement in reading, science, and mathematics of elementary and secondary students. It would make better sense to compare Wisconsin or Minnesota with Norway or Finland on these variables.

<sup>16</sup> The classic study is Robinson (1950). Cf, Holt., Steel, Tranmer, and Wrigley (1996), "Aggregation and Ecological Effects in Geographically Based Data". *Geographical Analysis*, Vol 28, no 3, pp244- 262.

for teachers, is similarly problematic because states differ dramatically in pass/fail standards (NAS study report). For example, a score that is far below the score required for passing in one state may be far above the score required in another state. Finally, graduation and retention rates are also not credible measures to compare across states on their own, because, again, they must be interpreted in context. A low retention rate may be purposeful at an institution dedicated to serving at risk populations. Does this mean that no state-based comparisons are possible? In fact states may conduct comparisons over time within their states to provide valuable benchmarking data about the quality of performance of graduates from their public institutions. Comparisons between states are also possible (See appendix A), States may also desire to establish minimum levels of performance outcomes for undergraduate graduates and benchmark them against the same measures in sets of states judged to be most similar to them.

## Assessment Principles

Using measures whose scores are interpretable across professors, colleges, and time allows for making relevant comparisons within and between institutions. For example, the scores on such measures can be used along with grades on other tests (such as the SAT or ACT) as controls to assess whether the students at a school are doing better or worse on an outcome measure than would be expected given their entry-level skills. Measures that are applicable across institutions also may serve as benchmarks for interpreting the results with similar but locally constructed instruments or course grades. Measures that are designed to permit comparisons across institutions thus provide a signal of academic performance (and therefore motivator for change). Such signaling can indicate whether faculty and administrators need to take a closer look at the resources, curriculum, pedagogy, and programmatic structure underlying undergraduate teaching and learning. In short, such measures may help colleges document the progress they are making in fostering student learning. The measures also may contribute to improving academic programs by providing institutions with baseline and outcome scores to help identify the effects on learning of programmatic and curricular changes.<sup>17</sup>

---

17

With at least two noteworthy exceptions (the Cooperative Institutional Research Program (CIRP) and the National Survey of Student Engagement NSSE), efforts at higher education assessment have focused on developing approaches and instruments that deal with individual courses or majors that are often diagnostic in nature. However, one needs assessment data that permits comparison in order to successfully conduct formative assessment within institutions. The variation in learning goals



To accomplish these ends, cross-institutional measures must have certain essential characteristics. The scores must be reliable in the sense that they are not overly affected by chance factors. If the results are aggregated to the college level (such as to providing information about programs), then the degree of reliability required to identify effects is much less than would be needed for making decisions about individual students. The scores must be valid in the sense of providing information about student characteristics that are important to the institution's goals, such as improving their students' ability to communicate in writing and to think critically about issues.<sup>18</sup>

The process of implementing such measures at the college level is fraught with land mines. For instance, any top-down effort to impose them on faculty and students is likely to run into trouble.<sup>19</sup> Instead, it will be essential for the academic community to see them as a valuable adjunct to their own measures or even embed them into their own capstone courses. Similarly, attempts to use the results to punish institutions for having less than stellar or even average improvement scores would stop the assessment effort in its tracks. Instead, the results need to be used to identify best practices that other institutions could try as well as spot potential problem areas where additional support is needed

It is not feasible to measure all or even most of the knowledge, skills, and abilities that are central to a college's learning goals. Much of what is learned takes place outside the classroom. This situation leads

---

and teaching approaches vary enormously across American higher education. How can one know how well an institution is faring unless one compares the performance of the institution with that of other institutions?

<sup>18</sup> The tests themselves must be fair to all takers, i.e., regardless of the students' demographic or other background characteristics. The amount of testing time per student and the total costs of implementing the assessment package must be realistic; e.g., probably not more than three hours of testing time per student. One way to keep testing time low and still cover multiple skills covered by critical thinking, analytic reasoning, and writing, is to use a technique called "matrix sampling." This strategy (which works well when the college is the unit of analysis) involves assigning different sets of measures to different sets of students. In addition, the measures themselves must be intrinsically interesting and engaging so that students will be motivated to participate in the assessment activities and to try their best to do well. And finally, results have to be reported to students and institutions promptly and in a way that is understandable to the recipients and facilitates decision-making.

<sup>19</sup> Norms ceding power to faculty on key issues remain powerful. These norms were developed over time in the wake of the construction of the modern American university during the last quarter of the 19<sup>th</sup> century which married the graduate, research functions to the undergraduate mission. That move created the basis for the professional development of the doctorate as the final degree for faculty along with the recognition that only those who received the Ph.D. in their field should make decisions about the issues noted above. Central administrators and boards of trustees may decide the size of the undergraduate enrollment, the resources that go to each department or college. However, the faculty rule on curriculum matters, including whether and how to assess its quality.

to the concern that what is tested will be overly emphasized in the institution's instructional programs. In short, some will say that the only abilities that count are the ones that are measured. This position is akin to saying "you shouldn't measure anything unless you can measure everything." This concern can be addressed by varying the types of measures used over time and by augmenting the measures that are used across institutions with local program specific instruments. To make this discussion more concrete we next present the assumptions, goals, methods and results for the Collegiate Learning Assessment (CLA), a new initiative we have been developing.

### The Collegiate Learning Assessment

We have been working with diverse colleges around the country to explore the feasibility of implementing the foregoing principles on a large scale. This research has involved the following testing activities:

- Tests are used to assess student skills (See appendix B for an illustrative measure).<sup>20</sup>
  - All the measures are open-ended; i.e., students write essays or short answers to the questions in each task. These measures fit within Shavelson and Huang's (2003) framework for conceptualizing, developing, and interpreting direct measures of student learning.
  - The measures are delivered to students over the Internet. The students take the tests under standardized exam conditions in their college's computer labs. This test administration procedure greatly reduces the costs of the assessment process and helps to insure data quality.
  - The students enter their answers online, and the responses are processed and scored by computer. Computer software programs "learn" how to score the open-ended responses based on the task's scoring rubric and a sample of 400 answers that were graded by human scorers.
  - A given student takes only a small portion of the entire set of tests administered at each college. This "matrix sampling" approach uses measures from several clusters of disciplines (e.g., natural
-

sciences, social sciences, and arts and humanities) but only requires two to three hours of testing time per student.<sup>21</sup>

- The students' SAT or ACT scores (which are obtained from their campus' registrar's office) are used to put the scores from the admissions tests on a common scale and to adjust for differences in admissions and grading standards across colleges.
- The "unit of analysis" is the college. Although data are collected on individual students, these data are aggregated to the institution, because that is the locus for program improvement. However, this focus does not preclude examining separate colleges or programs within a large institution.<sup>22</sup>
- Colleges are informed about the average of their students' scores (individual scores are only reported to the students). They also are advised about whether their average is above or below what would be expected given their students' mean SAT or ACT scores. Although anonymous institutional averages are presented (such that a college can compare itself to others in the sample), no college's identity is disclosed to any other institution.
- Research is conducted to assess the reliability and validity of the scores assigned, the relationship between these scores and other measures, the interaction between task type and student characteristics (including demographics and academic major), student motivation, the characteristics of the schools that have average scores that are above or below the expected level, and other factors.

Conducting the above-listed activities, a proof of concept study was conducted with 14 colleges and universities testing 1365 students in the 2002-2003 academic year. The tests were administered in computer labs where most students typed their answers on computer discs provided for the purpose based on instructions and questions provided on paper. Five findings emerged from our feasibility study (Klein, et al., 2004): the measures satisfy psychometric standards for reliability and validity; the graders agree highly with each other in the scores they assign to an answer; students exhibit consistent performance across tasks; and after controlling on their college admissions scores, seniors and juniors earn significantly higher scores on our tests than do freshmen and sophomores. This latter finding indicates the measures are sensitive to the amount of education a student receives (recognizing that learning occurs

---

<sup>21</sup> See Klein, et al. (2005) for description of matrix sampling.

<sup>22</sup> And as we will argue below the focus on the institution as the unit of analysis does not foreclose the possibility of doing comparisons across states. However, such comparisons need to be based on carefully developed ways to compare institutions in a state with institutions in other states (see appendix a).

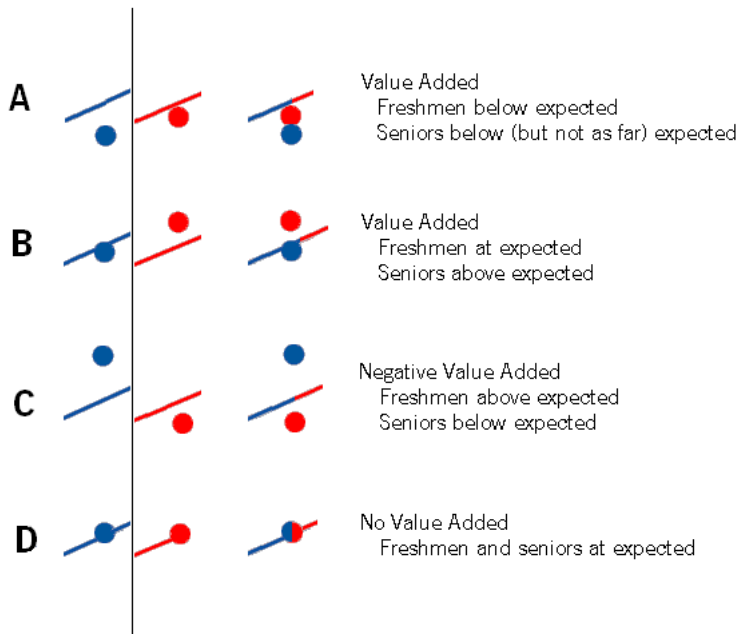
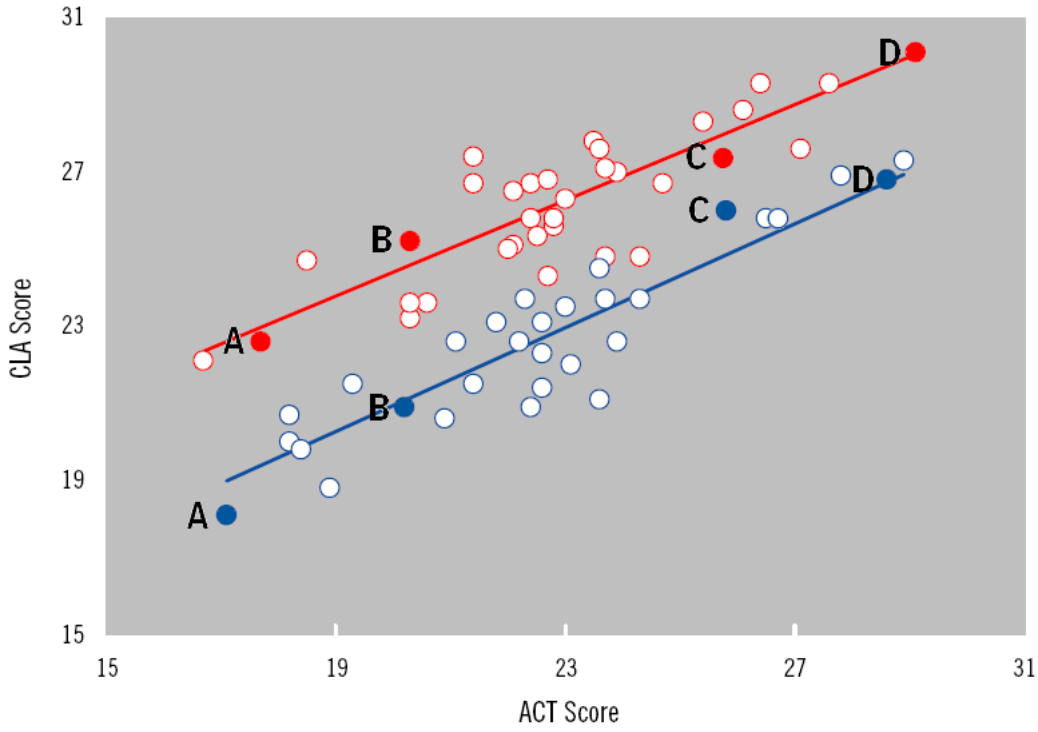
both in and out of the classroom).<sup>23</sup> In addition, the substantial correlation between these scores and the students' college GPAs suggests that the types of tests we are using measure abilities that are relevant to the educational process. Finally, we replicated the findings of others who have reported a high correlation between human and computer scoring of open-ended responses by comparing hand scoring of writing prompts with computer scoring done by the Educational Testing Service (ETS).<sup>24</sup> For example, we obtained a 0.95 correlation between these two scoring methods when the school is the unit of analysis.<sup>25</sup> The major advantages of computer scoring and Internet testing are the substantial savings in test administration, scoring and reporting costs, and elimination of the numerous problems associated with reading hand writing and handling hard copy, and much faster turnaround time in reporting results.<sup>26</sup>

The feasibility study further found that some colleges had average scores that were significantly higher or lower than would be expected on the basis of their mean SAT scores. Subsequent results, based on testing 45 institutions in the academic year 2004-2005 refine and extend the findings of value added within and between institutions. Figure 1 extends the finding of both within school value added growth and between institution differences. Moreover, Figure 1 shows that selection effects evident in the entering class are altered by institutions by the time students graduate. There is almost a mean 1.6 standard deviation growth in value added within the 45 institutions. And the figure shows some institutions improving significantly from the freshmen to senior level. Students said the tasks were engaging and the faculty found them to be authentic, i.e., realistic tasks that faculty think it is reasonable to expect graduating students to be able to perform.

---

<sup>23</sup> This is notable because previous longitudinal and cross-sectional studies that utilized multiple-choice indicators have not found such differences (Pascarella and Terenzini (2005). Still, an issue that faces all educational assessment is the difficulty in parsing out the educational contribution of a particular institution (as separate from general skill development and learning that could happen regardless of which college a student attends) or even learning that might happen if the student did not attend college (also called maturation effects). Further complicating this matter is that 60 percent of students attend more than one institution while pursuing their undergraduate educations. We will address these concerns in our sampling and analysis methods.

<sup>24</sup> ETS has also computer scored a sample of our performance tasks at about the same level of reliability and validity as human scorers.



## Figure 1

Value Added Comparison Within and Between Colleges and Universities<sup>27</sup>

The findings above, cost calculations, and other considerations indicate that a large-scale assessment system is feasible provided that (1) the measures are administered over the Internet, (2) open-ended responses are computer-scored, and (3) there are enough different tasks to draw upon to achieve reasonable test security and provide faculty and students with sample prompts.<sup>28</sup>

## Notes for Reconciliation

States are increasingly developing assessment systems that emphasize accountability. Resistance by faculty to accountability oriented systems of assessment (that are focused on indirect, proxy measures of student learning) also continues and is unlikely to change. This is an unfortunate, even problematic situation if, as we believe, the state level demand for accountability is only going to grow. We should reject the argument that the unit of analysis for accountability must be only the state<sup>29</sup> or the argument that the unit must only be the institution.<sup>30</sup> How might we reconcile the implications of the two units of analysis? We argue that the prime focus of accountability should be on student learning. And we will also argue below the focus on the institution as the unit of analysis does not foreclose the possibility of doing comparisons across states. However, such comparisons need to be based on carefully developed ways to compare institutions in a state with institutions in other states.

Representatives of the state and institutions, including both their administration and faculties, will need to work out the equivalent of a legal agreement that both parties will adhere to. In most cases the venue for this activity will be the state-based higher education coordinating commissions. These rules of engagement must give both parties incentives to cooperate. What should the rules of engagement be? First, there must be agreement on the measures to be used. The measures must meet faculty objectives but the ability for inter-institutional comparison should be built in to satisfy the needs of the state. Although the two parties need to agree on common measures to be used, their goals are different. Since

---

<sup>27</sup> Based on Figures 1-3, Collegiate Learning Assessment (CLA) Institutional Score Report 2004-2005, Council for Aid to Education, New York, NY, p. 9.

<sup>28</sup> We have continued to refine our administration of the CLA over the Internet. In the current academic year 2005-2006 we are testing 40,000 students in 120 colleges and universities.

<sup>29</sup> Callan et al. (2000; 2001; 2003).

<sup>30</sup> Banta et al. (1996).

faculty are primarily interested in assessment for educational improvement objectives while the state is primarily interested in assessment for accountability goals, the two parties will need to reach agreement on what information from the assessments may be aggregated at the regional or state level.<sup>31</sup> Relations between the institutions and the state will be considerably improved if there is agreement that the focus should be on improvement in the value-added contribution of the institution to student learning over time rather than a focus on absolute levels achieved. Indeed, if there is agreement that the value-added approach is appropriate, there can be a time lag built in during which institutions identified as being below minimum levels of quality can be asked to show improvement over a several year period. Since institutions, as well as the state, are interested in demonstrating that they are improving, this strategy should provide common ground between the two groups. Eventually, parties using the CLA will also want to establish criterion referenced norms to define a reasonable minimum standard for CLA performance. Currently, for example, each state defines and, if it so desires, changes its own cutoff points for satisfactory performance in high stakes testing in K-12 education. The National Assessment of Educational Progress (NAEP), developed by education testing experts, at least offers an alternative way to think about these norms. Thus it may be more appropriate that higher education testing experts take the lead in establishing cut off levels for satisfactory to excellent performance on the CLA. Clearly having comparative data is a necessary condition for helping to decide what a fair standard should be. The CLA will <sup>32</sup>help with the discussion of what that standard can/should look like once we see the span of scores and modality of scores from a few hundred institutions that span the Carnegie classifications having taken the CLA.

Governance of this partnership will also need to be considered carefully. Ideally, an independent commission might be set up to govern the relationship between the state and the institutions. This is unlikely to occur. What is more likely is that existing higher education coordinating commissions will be given the responsibility to implement any agreement to assess higher education institutions within their states. What is most important under these circumstances is that the agreements be carried out faithfully and consistently, within the terms of the rules of engagement. Anything else will lead to breakdown between the institution and the state.

---

<sup>31</sup> On the relationship between assessment focused on the formative (the institutional focus) versus summative (the state) the literature on this subject developed with respect to K-12 assessment is instructive. See William, and Black (1996) and Shepard, (2003); See also Shavelson, et al. (2004)

<sup>32</sup> Two hundred institutions have successfully tested thus far. At least that many will test in 2006-07 through 2007-08.

## Conclusion

There is a disconnect in assessment and accountability goals focused on student learning between the institution and the state. Can it be overcome by the state exerting control through its levers of power, i.e., the power of the purse or regulation? Probably not or, to put it another way, the result would certainly be a pyrrhic victory with no winners on either side. Can the disconnect be bridged? The answer is yes. It appears that, increasingly, state leaders will be judged on how well they improve the skills of their workforce to make their states more competitive economically.<sup>33</sup> If they do not succeed in doing so, they will not be successful in raising the nature of their state-based economy up the curve of valued added economic activity which, in turn, will mean the best educated members of their workforce will leave. Faculty and administrators should come to recognize the right of state political leaders to be concerned about the quality of undergraduate education and therefore have the right to set goals for improvement in student learning outcomes at higher education institutions in their state. Education is the main venue to accomplish this goal. Therefore we can expect heightened attention by state leaders on the performance of their higher education institutions as well as their K-12 system. Eventually, along with the growing recognition that the role of the state in setting goals is reasonable, should also flow state-based incentives, accepted by higher education leaders as appropriate, to encourage their public higher education institutions to meet those goals.<sup>34</sup> This is so because of the growing recognition, by all parties, that human capital is the most important asset a region, state, or nation has.<sup>35</sup> However, in the case of higher education, reliance on the experts (the faculty) to define the most appropriate methods of assessment is, necessarily, a prerequisite to success. This recognition of the need to work together by faculty and administrators at colleges, on the one hand, and state leaders, on the other, may well take some time and the road getting there will likely be bumpy. However, if human capital is as important as we believe, state and national leaders will ultimately be entrusted with the task of setting standards for improvement in student learning. If they do not, the consequences in a globally competitive economic environment will be severe. However, if we reach a wider consensus on how to implement this principle, we will be able to

---

<sup>33</sup> For example, see the publications devoted to aspects of this topic on the National Governors' Association (NGA) website. The NGA is the official association for the fifty governors.

<sup>34</sup> How this process plays out and how long it will take is, of course, an important question. And whether sanctions will be employed by states is also an important question. Ideally, the process will involve extensive conciliation between institutional and state-based leaders.

<sup>35</sup> See Krueger (2003).



develop policies and practices in assessment that benefit the institution and the state and, most importantly, the citizens both serve.<sup>36</sup>

---

<sup>36</sup> The heart of the issue is how to arrive at decisions that are in the public interest. In viewing the higher education institution versus the state it is overly simplistic to argue that the state narrowly defined as its bureaucratic structure, alone, is, represents or speaks for the public interest. This is so because the concept public interest itself is a tenuous concept. Moreover, the university itself is seen as being a venue for the public interest (See Coady, 2000). The best outcome is a careful balance between the state and the institution with both sides given rights and responsibilities here.

## Appendix A

### A Strategy for State-Based Comparisons

We suggest the following strategy for within and across state comparisons. Instead of making direct comparisons among states on such measures as graduation and retention rates, passage rates on licensing exams for teachers or nurses,<sup>37</sup> the logic of the CLA approach suggests that one make comparisons among institutions grouped by states. If one simply aggregates the scores of all institutions within a state to create a single state score, this eliminates the ability to understand the range of variation of the value added scores of specific institutions which, in turn, could seriously skew the results. . Second, instead of making comparisons in the level of student performance between states, the CLA method calls for comparing the value-added scores of colleges within and between states. Why look at value added compared to absolute level of student learning outcome levels? The answer is that while absolute levels achieved are interesting to note, the CLA approach places the focus of attention on improvement. We believe this is a more realistic and fruitful strategy to pursue. In addition, after a period of time in which the value added growth of a state's colleges are benchmarked, goals for value added improvement for the colleges may be established. And, eventually, state and college leaders across states and/or within a state may work to establish minimum levels of student performance on the CLA instruments.

To do what the CLA requires, we always need an input measure that is applicable across all the institutions in the study (mean SAT or ACT scores or a measure correlated with SAT or ACT scores for all students taking the CLA tests).

The CLA approach can be used to make comparisons between states by computing the proportion of a state's schools and/or students that are well above, above, on, below, or well below a plotted regression line. This comparison answers the question of how effective are a state's colleges in improving student performance on the outcome measures assessed relative to the effectiveness of the colleges in other states.

Here are three graphs that illustrate this approach. First, Graph One shows the scores of colleges in three states plotted on a regression line. When we break out the scores of the colleges by state (Graphs Two and Three) we indeed find that one state's colleges are doing considerably better than another state's

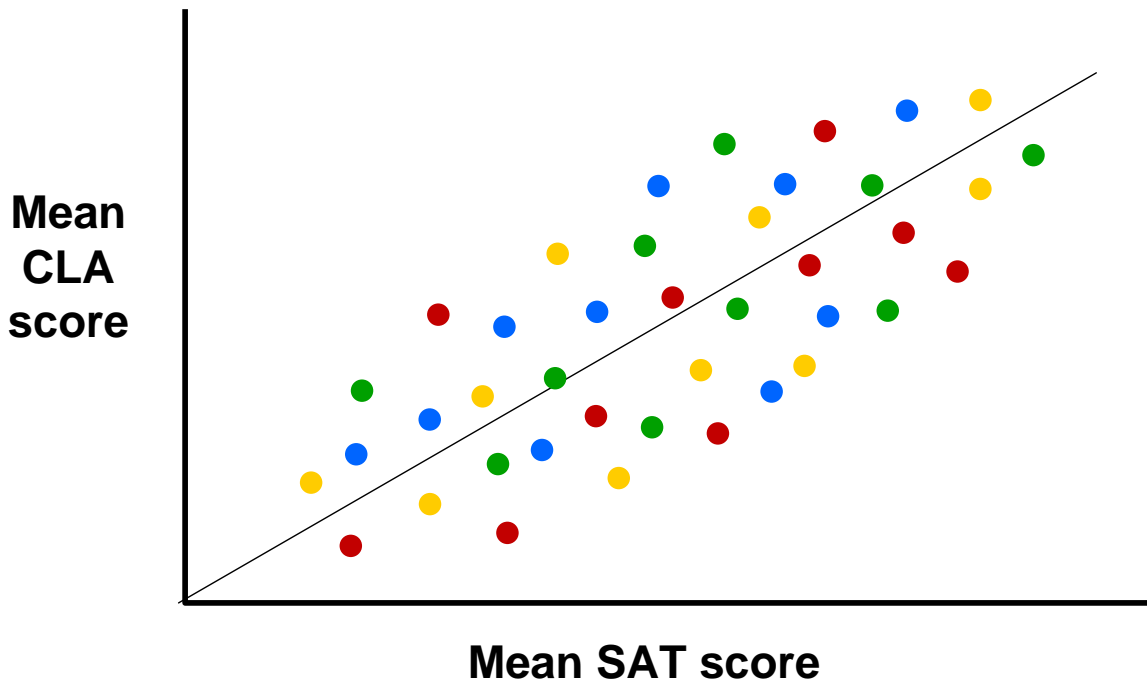
---

<sup>37</sup> See Callan, et al. (2004).

colleges. The question then becomes why which leads to examining the best practices of those institutions that are doing well.

Graph One  
Scores of Colleges in Three States

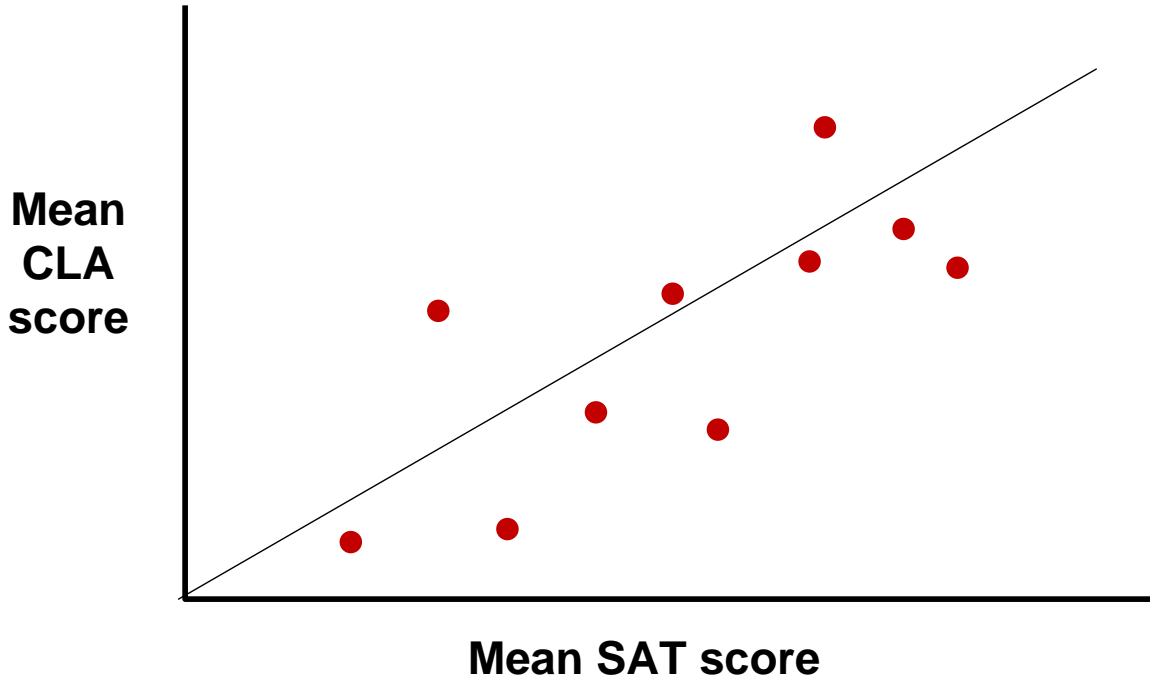
## ***Example of Institutional and State Comparisons***



Graph Two

Most Value Added Gains for State A are Less than Expected

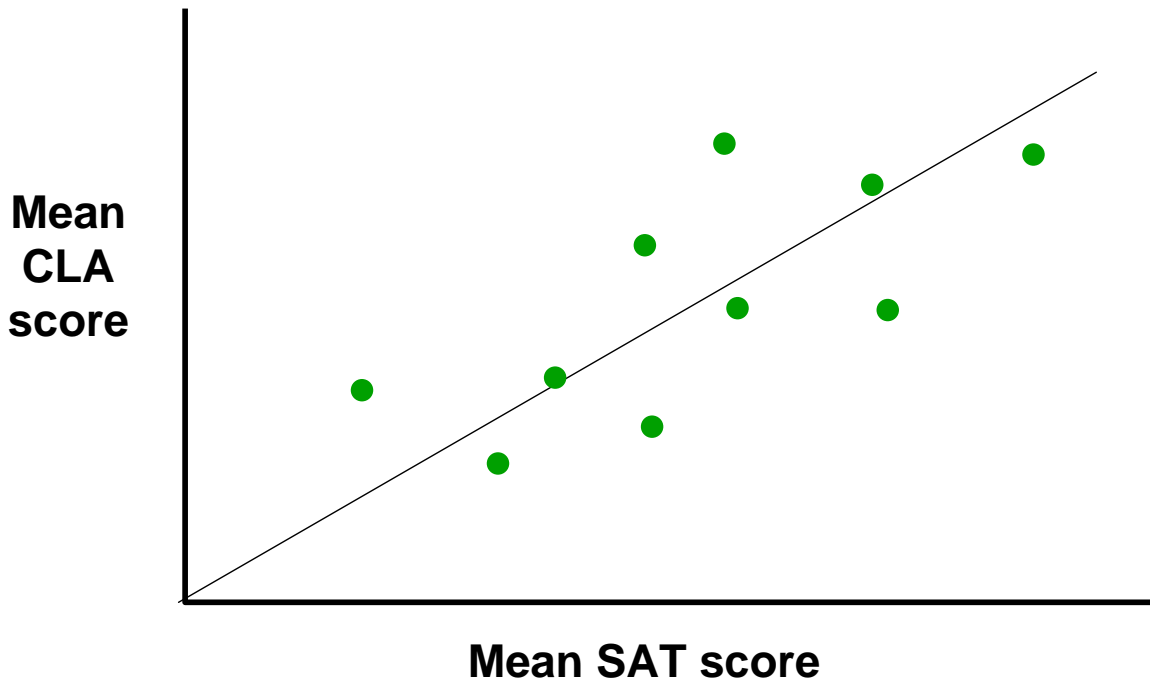
***Example of Institutional and State Comparisons***



Graph Three

State B Shows Mixed Value Added Gains

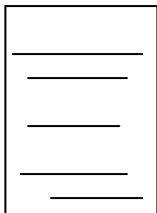
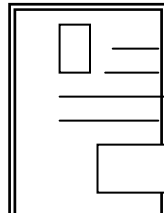
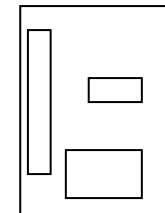
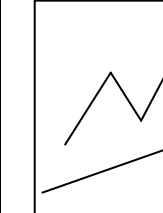
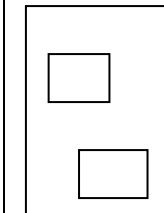
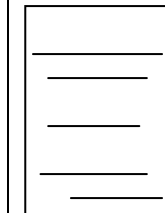
## ***Example of Institutional and State Comparisons***



## Appendix B

Illustrative CLA Performance Measure<sup>38</sup>

You are the assistant to Pat Williams, the president of DynaTech (which is a company that makes precision electronic instruments and navigational equipment). Sally Evans, a member of DynaTech's sales force, recommended that DynaTech buy a small private plane (a SwiftAir 235) that she and other members of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation:

					
Pat's e-mail to you and Sally's e-mail to Pat.	NTSB report on in-flight break-ups	Newspaper articles about the accident	Airplane characteristics chart.	Amateur Pilot article (featuring SwiftAir's customer training policy)	Pictures and description of SwiftAir Models 180 and 235

You are asked to prepare a memo in which you address several questions, including what data support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups, what other factors might have contributed to the accident and should be taken into account, and your overall recommendation about whether or not DynaTech should purchase the plane.

Students are given 90 minutes to complete the task. The scoring rubric assigns points for a number of factors. There also are overall scores for analytic reasoning and communication skills.

## References

- Association of American Colleges and Universities (AAC&U) (2002). *Greater Expectations: A New Vision for Learning As A Nation Goes to College*. Washington D. C.: AAC&U.
- Banta, T. W., J. P. Lund and F. W. Oblander, (eds.) (1996). *Assessment in Practice: Putting Principles to Work on College Campuses*. San Francisco: Jossey-Bass.
- Burke, J. C. and H. Minassians (2002). *Performance Reporting: The Preferred “No Cost” Accountability Program* (2001). Albany: The Nelson A. Rockefeller Institute of Government.
- Callan, P. M., et al. (2000). *Measuring Up The State Report Card*. San Jose, CA: The National Center for Public Policy and Higher Education.
- Callan, P. M., W. Doyle, and J. E. Finney (2001). “Evaluating State Higher Education Performance,” *Change*, March/April, pp. 10-19.
- Callan, P. M. and J. E. Finney (2003). *Multiple Pathways and State Policy: Toward Education and Training Beyond High School*. Boston, MA: Jobs For the Future.
- CLA Institutional Score Report, 2004-2005 (2005). New York, NY: Council for Aid to Education.
- Coady, T. (ed) (2000). *Why Universities Matter: A Conversation about Values, Means, and Directions*. St. Leonards, Australia: Allen & Unwin Pty.LTD.
- Collier, D. (1991). “The Comparative Method: Two Decades of Change,” in D. A. Rustow and K. P. Erickson (eds.), *Comparative Political Dynamics: Global Research Perspectives*. New York, NY. Harper Collins.
- Dondio, R. (2005). “The Tempest in the Ivory Tower,” *Book Review*, New York Times. March 27.

- Ewell, P. T. (2002). "Grading Student Learning: You Have to Start Somewhere," in *Measuring Up 2002: The State-by-State Report Card for Higher Education*. San Jose, CA: The National Center for Public Policy and Higher Education, pp. 73-76.
- Holt, D., D. G. Steel, M. Tranmer, and N. Wrigley (1996). "Aggregation and Ecological Effects I in Geographically Based Data". *Geographical Analysis*, Vol 28, no 3, pp 244- 262.
- Jones, D. (2003). *State Shortfalls Projected Throughout the Decade*. National Center for Public Policy in Higher Education. San Jose, CA.
- Immerwahr, J. (August, 2000). *Great Expectations: How the Public and Parents—White, African-American and Hispanic View Higher Education*. San Jose, CA: National Center for Higher Education Public Policy.
- Immerwahr, J. (2002) *Meeting the Competition: College and University Presidents, Faculty, and State Legislators View the New Competitive Academic Arena, The Futures Project: Policy for Higher Education in A Changing World*. Providence, R.I.: Brown University.
- Immerwahr, J. (2003). *Public Attitudes on Higher Education: A Trend Analysis*. San Jose, CA: National Center for Higher Education Policy.
- Klein, S., Kuh, G., Chun, M., Hamilton, L., & Shavelson, R. (2005). An approach to measuring cognitive outcomes across higher-education institutions. *Journal of Higher Education*, 46, No. 3, 251-276.
- Krueger, A. B. (2003). *Education Matters: A Selection of Essays on Education*. London: Edward Elgar.



Measuring Up (2002): The State-By-State Report Card for Higher Education. San Jose, CA: National Center for Public Policy in Higher Education.

Measuring Up (2004): The National Report Card on Higher Education. San Jose, CA: National Center for Public Policy in Higher Education.

Naughton, B.A., A. Y. Suen, and R. J. Shavelson (2003). "Accountability for What? Understanding the Learning Objectives in State Higher Education Accountability Programs." A paper presented at the annual meeting of the American Educational Research Association. Chicago.

Pascarella, E. and P. Terenzini. (2005). How College Affects Students: A Third Decade of Research. Jossey-Bass.

Ragin, C. C. (1989). The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies. Berkeley, CA: University of California Press.

Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals, American Sociological Review, Vol. 15, pp. 351-357.

Shavelson, R.J., & L. Huang. (2003). Responding responsibly to the frenzy to assess learning in higher education, Change, 35(1), 10-19.

Shavelson, R. J., P. J Black, W. Dylan, and J. Coffey (2004). On Linking Formative and Summative Functions In The Design of Large-Scale Assessment Systems, submitted to Educational Evaluation and Policy Analysis.

Shepard, L.A. (2003). Reconsidering large-scale assessment to heighten its relevance to learning, In J.M. Atkin & J.E. Coffey (Eds.), "Everyday assessment in the science classroom." Arlington, VA: NSTA Press.

William, D., & P. Black (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, 22(5), 537-548.