UNC System Faculty Assembly

Report:

Characteristics of “Best Assessment Instruments”

and Best Practices for Student Assessment

Student Assessment Subgroup of the

Assessment Task Force

23 October 2006

Executive Summary

Universities and university systems should adopt instruments to assess student learning or experience when they have clear understandings of what sorts of data those instruments will provide and specific plans for using that information to improve education. Assessment instruments are less valuable for providing data to the public, because in order to be meaningful, statistical data on student achievement have to be contextualized in ways that can overcome their utility as simple comparative metrics. Differences in institutional mission, resources, and peer group, in the populations and constituencies individual universities must serve, and in student background affect the interpretation of comparative data on student learning in important ways. In addition, the number and complexity of educational goals at the post-secondary level makes any given metric of student learning incomplete and potentially misleading. Therefore, learning assessment instruments are best approached as tools for identifying potential problems, while decisions about how to improve outcomes must rely on much broader sets of data generated by different means and subject to problem-solving discussions by multiple campus constituencies.

Characteristics of Best Assessment Instruments. The best learning assessment instruments provide kinds of data substantially different from those already available to faculty and administrators. These data are agreed to be of utility to both teaching faculty and administrators, and are consistent with learning outcomes already mandated by accreditation agencies. They do not impinge unreasonably on instructional time, personnel, or resources, and do not influence instructors to "teach to the test." They measure skills agreed by faculty to be central goals of college education, and while tailored carefully to different institutional missions, they allow specific, meaningful, and actionable comparisons between institutions. They provide information specific enough to evaluate the success of particular programs and curricula, and

their vendors provide institutions with at least some access to their own data for the purposes of further research and analysis. Their design and statistical procedures differentiate explicitly between the effects of student background, aging and maturation, institutional structure and environment, and the effects of an institution's curriculum in contributing to student learning. They protect the privacy of student information and can be demonstrated to be free of ethnic, class, and gender bias.

Best Institutional Practices for Adopting and Using Assessment Instruments. Before adopting assessment instruments, institutions should articulate persuasively and successfully defend their rationale, outlining clearly the steps to be taken in collecting, analyzing and applying the results. They should specify how and by whom results will be evaluated and what resources will be available to correct perceived shortcomings. Results should be approached critically. They should not be used as direct guides for policymaking, but should help identify issues to be subjected to independent scrutiny and checked against different sorts of data. Instrument administration and the development of new policies and practices in response to assessment data should neither draw resources from classroom instruction, nor threaten withdrawal of resources from "low performing" institutions or units. Assessment instruments should not be used as exit examinations.

NSSE and CLA. The National Survey of Student Engagement (NSSE), despite its problematic dependence on self-report surveys, provides a rich variety of information institutions can use to understand student time investment and other elements of their college experience. It appears easily tailored to the concerns of institutions with different missions and student populations. The Collegiate Learning Assessment (CLA) uses innovative methods to evaluate student critical thinking (CT) skills, but correlates so highly with SAT scores that it may provide

little new information. CLA does not yet have longitudinal data with which to evaluate the progress of individual samples of students, and does not yet compare scores by institutional peer groups. Problems with its sampling, statistical analysis, and data reporting hamper CLA's ability to provide universities with unique, reliable data on the success of their curricula in improving CT skills

> It is counterproductive to make decisions based on assumptions derived from unexamined numbers. Yet that is what we in higher education do when we fail to question statistical assertions, when we fail to triangulate–that is, to find other sources and types of evidence to affirm or contradict those assertions. We have been gulled by a propaganda of numbers that has shaped how we think about the enterprise. It is our responsibility to exercise due diligence in generating, interpreting, and responding to statistical assertions, particularly those from unofficial sources. If we don't the propaganda of numbers will turn into tyranny.
>
> Clifford Adelman, "The Propaganda of Numbers,"
> *Chronicle of Higher Education* 13 October 2006, p. B6

Composition

The Student Learning Subgroup of the Assessment Task Force of the University of North Carolina Faculty Assembly was convened on 29 September 2006 and given a deadline of 27 October to formulate this report. The Student Learning Subgroup consisted of Gwen McNeill Ashburn from UNCA, Andrew Koch from ASU, Meg Morgan from UNCC, LeRoy Percy from NCSA, Robert Mark Spaulding from UNCW, and Gregory Starrett from UNCC. Starrett was elected Chair of the group.

Charge

The Student Learning Subgroup was charged with identifying "best practices for assuring quality measurement while protecting student privacy and minimizing disruption [of classroom learning time]." More specifically, we were charged with "establishing clear criteria that should be used in evaluating" the National Survey of Student Engagement (NSSE) and the Collegiate Learning Assessment (CLA), two evaluation instruments developed and marketed by private educational

testing companies in cooperation with a number of non-profit foundations and research institutes. Both instruments are currently in use, or have been used, by various NC System campuses.

Action

During its 29 September meeting the Subgroup discussed information made available to us on NSSE and CLA. The Chair recorded the main points of this discussion and has since gathered supplementary information on these two instruments, particularly CLA. A draft of this report was circulated by e-mail to Subgroup members on 16 October 2006 with a request for comments for revision.

Background and General Considerations

Assessing student experience in college cannot be approached as a narrowly technical enterprise. The choices leading students to particular colleges and the benefits they derive from their experiences there are complex. Decisions to adopt evaluation instruments for assessing student learning and other elements of student experience at college need to be made with this understanding, as well as with an appreciation of the multiple and sometimes conflicting demands made of postsecondary education by different constituencies. These include accrediting agencies, students, parents, alumni, legislatures, and numerous groups within civil society: schools, hospitals, businesses, churches, government agencies, and countless others who depend on the creativity and labor of our graduates. The public utility and value of higher education is not defined solely by specific learning outcomes. For example, individuals with higher education report greater levels of health, are less likely to use various forms of public

assistance, vote at substantially higher rates than those with a high school education or less, and are more actively engaged in volunteering (Baldwin and Pasque 2006). Likewise, the personal utility and value of higher education are not defined solely by the acquisition of specific quantifiable skills, but by the development of social and intellectual maturity, by the establishment of social networks, and by the appreciation of diversity and complexity in the world. As Clive Crook, a senior editor for *The Economist* and *The Atlantic Monthly* has written, responding to the precipitous narrowing of discourse on the value of higher education, "[E]nlightenment, not productivity, is the chief social justification for four years at college" (Crook 2006:28).

No single evaluation instrument or combination of instruments can capture all the information we might want in order to assess and help us maintain and improve the quality of higher education. Some elements of student learning and experience can be assessed quantitatively and supplement information we already have about the ingredients of student success. Other elements of student experience which influence learning are better approached with other methods (e.g. Nathan 2005, Holland and Eisenhart 1990).

Calls for new forms of "accountability" and assessment are generally made in environments in which people perceive real or imagined resource shortages, perceive real or imagined mismanagement of resources, or come to doubt the suitability of practices previously taken for granted. Educational institutions are designed to change, and particularly since the Second World War have adapted continuously and progressively to rapidly changing technologies, demographics, and social and political demands. They have done this based on a number of internal administrative decision-making and planning processes, with the help of external accreditation agencies which monitor the quality of services provided by colleges, and

in particular in response to faculty responsiveness to new educational practices and techniques, new directions in research, and changes in student interest and demand.

The professional assessments that faculty make of each other, the assessments students make of faculty, and the assessments administrators make of institutional goals and processes are not new. What is new is the claim that, despite their historical effectiveness and responsiveness, universities and colleges are "unaccountable," that they are averse to change, and that the quality of their contribution to the public good is in decline. None of these charges bears scrutiny. But for a number of reasons, many of them outside the control of universities and colleges, calls for "true" assessments of student learning have proliferated over the last few years, culminating most recently in calls by Secretary of Education Margaret Spellings for public reporting of data on student learning outcomes. Although the charge of this Task Force is a response to longer-standing issues internal to the UNC system and is not a direct response to the Spellings Report, the charge must now be read in the light of that report.

Assumptions

Current proposals to assess student learning often focus on the notion of "critical thinking." Critical thinking is a complex concept. Like the notion of "intelligence," it can have multiple components and be used differently in different domains of life. Habits of critical thought applied to mathematical problem-solving are not necessarily transferred or transferrable by individuals to artistic production, to political beliefs, or to religious participation. Likewise, the habits of critical thought that classicists apply to textual editing and reconstruction do not necessarily transfer to the ability to solve differential equations or understand currency markets. In everyday social interaction, critical thinking often makes things more rather than less

complicated. It is often inconvenient. And because of this it is sometimes unwelcome in large organizations and in the broader public sphere. The extent to which "critical thinking" as well as "accountability" are fads and buzzwords in the current educational and political environments must be borne in mind when making decisions about assessment.

Broader social goods like economic prosperity and technological development are not primarily the result of any particular set of skills or mode of thought taught in college, but have to do with the structure of labor markets, interest rates, the cost of energy resources, class stratification, private and public investment, and dozens, if not hundreds, of other factors (Crook 2006). General critical thinking skills–as opposed to specific technical skills in chemical engineering, software design, or financial management--are socially, politically, and morally important, but not necessarily the decisive factors in economic or technological progress.

Furthermore, the issue of serving the public by assessing and reporting student learning outcomes in order to make college comparisons easier is also more problematic than it is made to appear by higher education's critics. Parents and students do not evaluate the attractiveness of particular colleges solely in terms of learning outcome, but on the basis of selectivity, cost, location, reputation, unique curricular features, family connections, and institutional resources for publicity, among many other factors. Students are not infinitely mobile and often do not always have a choice between institutions. Therefore they cannot necessarily use data that might be provided about institutional success on learning outcome measures, and may end up feeling more disheartened than ever at the implication-by-numbers that their school may not be doing as well for them as some other school might have. The use of an idealized and oversimplified economic model in which people "shop around" for the best services is as misleading in higher education as it is in health care. "Accountability" through measuring and reporting student

learning outcomes to the public will only become meaningful for many families once the more troublesome issues of accessibility and affordability have been solved.

In the following three sections of this report, we summarize the characteristics of the best sorts of survey instruments for assessing student learning; the best practices for institutions in evaluating and making decisions about adopting assessment instruments; and some of the advantages and disadvantages of the two specific instruments mentioned in our charge. We did not attempt to rank the lists of best characteristics and best practices.

Characteristics of "Best Assessment Instruments"

The best assessment instruments:

- provide information of proven utility to both teaching faculty and to administrators.
- do not impinge on instructional or other classroom time, and do not presume, predispose or influence faculty to "teach to the test."
- use, or are at least consistent with the highly specific requirements for student learning outcomes (SLOs) already mandated by accreditation agencies.
- measure skills which are broadly agreed by teaching faculty to be the goals of a general education.
- make meaningful, specific, and actionable comparisons between institutions.
- provide significantly new kinds of information which are not otherwise available to faculty or administrators.
- provide the university with information that is tailored to institutional mission.

- provide campuses with some level of access to data from their institution that will allow them to conduct independent analyses of questions and populations of interest to them.
- provide information which can be used to identify and evaluate particular kinds of programs and curricula which lead to educational success as that success is defined by particular institutions.
- differentiate between effects of student background, the role of maturation, environmental effects (e.g. the composition of the student population, the location of the institution, etc.), and the specific effects of an institution's curriculum. (E.g., the instrument's designers and users must not automatically assume the college experience of MIT students in a major city like Boston and those of New Mexico Institute of Mining and Technology students at the relatively isolated campus in Socorro are similar, despite similarities in their curricula, educational goals, and the SAT scores of their students.)
- are those whose providers can demonstrate that they are free of gender, class, and cultural biases (i.e., that both men and women, and members of different socioeconomic and ethnic groups perform on these tests at similar levels given similar levels of academic preparation).
- can draw fine distinctions between institution types and structures. They do not ignore issues of class size, the frequency of instruction by graduate students or part-time faculty (see Wilson 2006), and other elements of academic practice when gathering or evaluating the results of student learning assessments.
- safeguard the privacy of individual student results.

Best Practices for the Adoption and Use of Assessment Instruments:

Institutions best ensure the quality of student learning assessment when:

- Decisions to adopt particular assessment instruments are based on clear institutional needs for specific types of data rather than on the mere availability, popularity, or perceived political pressure for adopting those instruments.
- *Before* adopting particular assessment instruments, the institution is able to outline clearly and defend a rationale for the use of the instrument, a justification for its collection schedule (Every year? Every five years? When needed? Just this once? *Why?*), and a set of criteria that will guide members of the institution in interpreting its results. For example, if measures of student performance are judged by the instrument's provider to be "at expected level" relative to a peer group or other population, is this to be taken as an indication of success, or of failure to have done better than other institutions within the group?
- *Before* adopting particular assessment instruments, campus and System leaders identify explicitly what processes will be used and what resources will realistically be available to correct perceived deficiencies, if any, identified by these assessment instruments. Otherwise institutions are placed in the position of patients for whom genetic tests can predict the likelihood of specific forms of cancer for which there is no effective treatment. Is it worth reporting problems for which there is no practicable cure? Is it responsible to students, alumni, parents, trustees, faculty, communities, and other constituencies, to report bad news without having effective means with which to

overcome perceived shortcomings?

- Instruments are not used as direct guides for policy decisions, but are used to raise questions, identify issues, problems, and areas of improvement which can be subjected to further independent scrutiny using different methods.
- It is understood that results reporting correlations often do not point in straightforward ways to reliable inferences about causation, prognosis, or treatment. Thus the results of assessment instruments should be shared broadly within the campus community and used as the focus of problem-solving discussions between multiple constituencies, rather than remaining the possession or responsibility of any one.
- Instruments are treated as part of ongoing educational research on campuses, making use of the expertise of faculty in schools of education and in the social sciences, particularly experts in the sociology of education, who are familiar with the most sophisticated current research and statistical techniques, and can best help campus leaders contextualize, evaluate and interpret statistical procedures and results.
- Results are approached critically. Some subgroup members pointed to wide disparities between the results of NSSE surveys and those of corresponding FSSE (Faculty Survey of Student Engagement) surveys on the same campus. Faculty and students, for example, have broadly different ideas about what constitutes “critical thinking,” and whether or not their coursework developed or drew on their critical thinking skills. This sort of disparity demands both further refinement of the measures as well as careful attention to multiple possible interpretations to which the survey results lend themselves on particular campuses and system-wide.
- The adoption of assessment instruments, and the subsequent investment of institutional

time and effort in interpreting and acting on assessment data, do not draw on financial or human resources that might ordinarily be devoted to classroom instruction.

- Assessments are performed on samples of student populations for the purposes of improving instructional programs, and never used as exit examinations for individual students or broader student populations.
- Assessment data are not used as the basis for threatening or depriving specific "low-performing" institutions of resources.
- It is understood that providing the public with simple quantitative data purporting to represent student learning outcomes may not automatically be a public service. Such data can be misleading without extensive contextualization regarding institutional type and mission.
- It is understood that providing the public with simple quantitative data regarding student learning outcomes or other measures of "institutional effectiveness" may hinder rather than facilitate program improvement over the long run because of the possibility that institutional type and mission may become stereotypes of "success" and "failure" generally. Thus the fact that UNC Chapel Hill's retention rate is 26 percent higher than East Carolina's, and that its four-year graduation rates are almost three times as high, does not indicate that students at Chapel Hill are receiving an education somewhere between twenty-five and three hundred percent better. It indicates that UNCCH's student population is far different from that at ECU. The higher education field's definition of "peer institutions" and differentiations in institutional mission are not transparent and not necessarily meaningful to parents and students.

Specific Comments on Assessment Instruments

NSSE—The National Survey of Student Engagement

The goal of the NSSE is to collect information on "student participation in programs and activities that institutions provide for their learning and personal development." It provides information about how students spend their time and reflects "empirically confirmed 'good practices' in undergraduate education. . .reflect[ing] behaviors by students and institutions that are associated with desired outcomes of college" (NSSE website).

Advantages of NSSE

- NSSE is well established, non-intrusive, and provides campuses with enough data to differentiate between the experiences of different kinds of students (e.g., commuters vs. residential students). It seeks to measure an array of demands on student time and a number of important elements of collegiate experience and perception that can have implications for both academic affairs and student affairs professionals.

Disadvantages of NSSE

- One of the major shortcomings of the NSSE instrument is its exclusive reliance on self-report surveys for its collection of information on how students use their time. That self-reported activity is sometimes the least expensive means of data gathering does not make it the most accurate. It should be noted that there are alternative methodologies for conducting time and activity studies.

CLA—The Collegiate Learning Assessment

The Collegiate Learning Assessment's goals are to test students' "performance on tasks that require them to think critically, reason analytically, solve realistic problems, and write clearly" (CLA website). It is a three-hour test of critical thinking (CT) skills consisting of carefully constructed problem-solving exercises and writing assignments that are scored, respectively, by trained human graders and by computer software licensed from the Educational Testing Service.

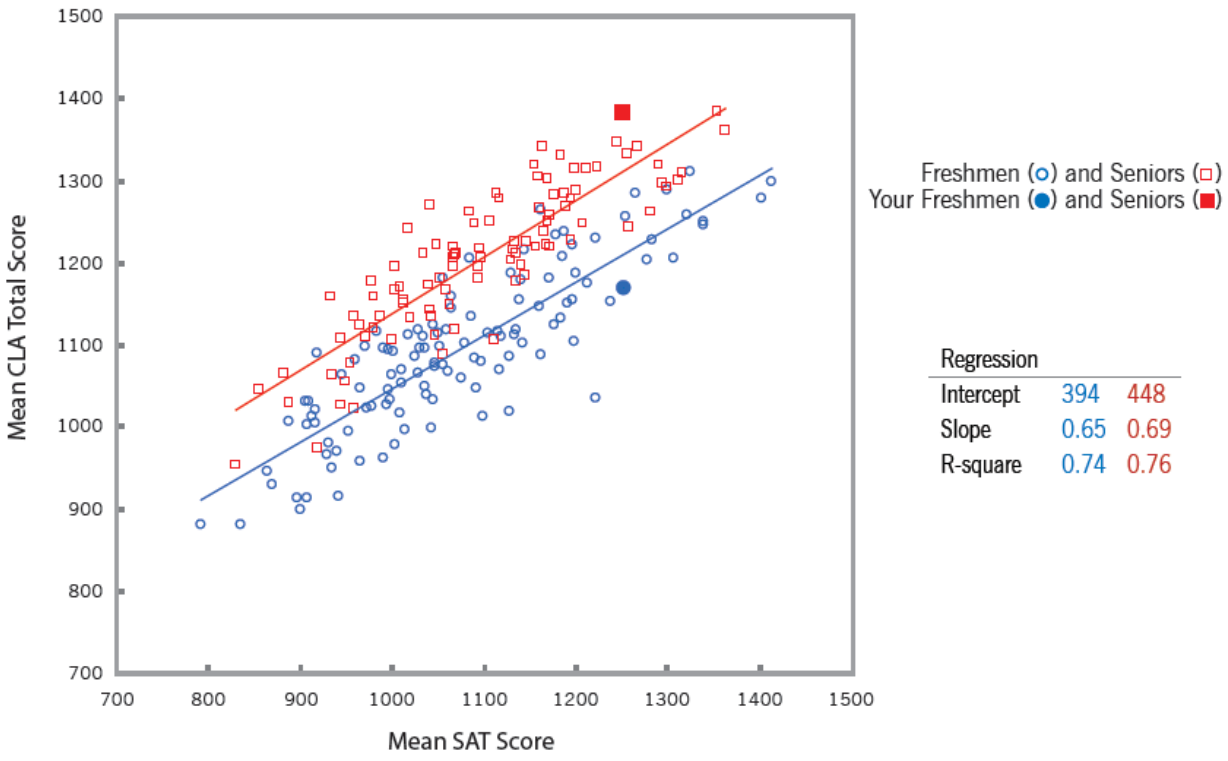
It is, as it claims, different from many other standardized tests in not relying on multiple-choice questions. In terms of measuring the intellectual skills normally associated with a college education, scores on CLA have extremely high correlations with ACT and SAT scores. CLA's data report r-squared values of between .75 and .80, meaning that seventy-five to eighty percent of the variance in CLA scores can be explained or predicted by variance in student SAT scores. So although the methodologies of the assessment instruments differ, it is clear that the intellectual skills required by the CLA are either extremely similar to, or co-vary directly with those required by the SAT. They test for similar sorts of skills in different ways. In fact, CLA uses SAT scores to construct "expected" values for CLA scores for both freshmen and seniors.

CLA offers, or plans to offer, two services to institutions: 1) a cross-sectional survey of CT skills derived from test data on samples of freshmen and of seniors in the same year; and 2) a four-year longitudinal study of the same sample population of students as they move from freshmen through the ranks of rising juniors, to seniors. CLA results were first collected in 2005, so there are currently no longitudinal data available. The first longitudinal studies are now under way.

The information provided to campuses by CLA consists of a global comparison of

institutional scores of all institutions who have administered the test to freshmen and to seniors, with the client campus's aggregate scores. This generally demonstrates a (much to be expected) difference between the scores of the sample of freshmen and the separate sample of seniors, graphed against their respective SAT scores (see Figure 1, below).

Figure 1: Relationship Between CLA Performance and Incoming Academic Ability



(Figure 1: From a sample "CLA Institutional Report")

CLA also calculates expected and real first-year retention rates, and 4- and 6-year graduation rates for the client campus, and presents disaggregated scores for student performance on the different sorts of tasks (problem-solving exercises versus analytical writing).

Advantages of CLA

The main utility of CLA is that it presents evidence of how one's own campus compares with other campuses in the test results of freshmen and seniors on specific sorts of CT exercises. It should be remembered that the skills evaluated by CLA are quite narrowly constructed. One of the advantages of CLA is that it is *not* a content-based test, whose administration would presuppose a rigid curricular uniformity across institutions.

On the other hand, because CLA results cannot be used as a shorthand or proxy evaluation of many of the most important skills and bodies of knowledge a student might bring away from college, it is by itself an inadequate measure either of the intensive labor expended by faculty, students, and staff at universities, or of the broader benefits of education to individual students. For example, CLA does not provide information about student preparation or success in STEM (Science, Technology, Engineering, Mathematics) fields; it does not provide information that will help evaluate student appreciation for historical or cultural heritage; it does not provide information that will help evaluate student understanding of diversity or student's skills in foreign languages. Each of these has been articulated as a vital goal for university education in the 21st century by different and often overlapping constituencies both inside and outside higher education.

Disadvantages and Difficulties of CLA

Sampling Issues

Some of these sampling and comparability issues may be resolved as more campuses participate in CLA testing in the future. Others are more deep-seated in the design and process of the CLA itself.

- Because its sample of participating institutions is not currently large enough, CLA does not segment its participating institutions into peer groups along any measure or set of measures. It does not report results by state, by institution type, or by other categorization of peer institution. The company plans to construct such categorizations in the future, but because the peer sets will be constructed inductively from data on participating institutions as the sample sizes grow, there is no way for institutions participating right now to know what their peer groups will look like.
- CLA does not compare test results to a control group of individuals who did not attend college, and so its results do not disentangle improvements due to age and maturation from those due specifically to participation in college or to the curricula of particular colleges. Nor does it control for other elements of student background or experience or institutional type which might influence test performance. There are statistical tests such as Hierarchical Interlinear Sampling (also called multilevel analysis) which can provide such differentiations, but at present CLA does not use or report them.
- Because CLA is currently only reporting cross-sectional data, which does not track specific groups of students over time, it may not capture significant changes in student populations. In addition, since seniors represent a group already subject to a number of filtering processes, the senior samples and the freshman samples in cross-sectional analyses are not comparable populations and differences in their scores do not necessarily measure curricular effects. This may also distort comparisons between schools with different retention rates.
- No data are supplied about the sorts of incentives provided for students at different institutions to participate in the project, meaning we have little evidence about how these

samples were identified and recruited. Good comparative data requires that samples be comparable across institutions not only in terms of degree type (students majoring in natural sciences versus humanities, for example), but in terms of other factors.

Comparability Issues

- CLA does not currently define or provide sets of peer institutions, but reports institutional scores in the context of all participating institutions, so the utility of its comparative data is thereby limited. Furthermore, once the longitudinal data do become available, users of CLA will need to wonder whether the automatic exclusion of transfer students from the sample, including all students entering with degrees from community colleges (who often do not enter as freshmen), will affect the validity of comparisons between campuses, whose rates of transfer student admission may vary significantly.

Data Reporting and Logistical Issues

- Because it reports data only at the institutional level, CLA does not help schools identify which of their programs are most effective in increasing student performance on these tasks. CLA's sampling protocol requires recruitment of students from five broad categories: Sciences & Engineering, Social Sciences, Humanities & Languages, Business, and Helping & Other. It does not, however, report to institutions on how students in these different fields fared on the exams, and its categorization of majors is sometimes bizarre, including both Law Enforcement and Visual and Performing Arts in the "Helping & Other" category, while unspecified "Multi/interdisciplinary studies" is placed in Helping & Other, and "Liberal/General Studies" is in the Humanities.

- CLA does not reveal the identity of schools which are performing “above” or “well above” expectations, resulting in a lack of models of success and the potential for a more or less random approach to identifying factors which might make a difference or strategies for improvement.
- Contrary to some claims, CLA administration is not “paperless,” at least not for university staff. It can be highly labor-intensive for campus employees, who must identify, recruit, track, motivate, and compile multiple consent forms and other records for hundreds of students. Finding ways to motivate students to take the demanding 3-hour tests can be difficult and costly, and different structures of recruitment and reward from campus to campus may affect the composition of samples and thus bias comparative use of the results. In addition, although campuses may be able to provide incentives for participating in the assessment, there may be no easy way to provide incentives for them to try very hard at it, since individual students are not required to release their scores to the institution. Pressuring students through reward or sanction to reveal their scores is ethically problematic.

Overall, despite its expense, CLA reports very little fresh information to institutions. In addition to measures and global comparisons of score differences between freshmen and seniors, it reports predicted and actual student retention rates. The latter are nearly identical and thus redundant to retention information already available from other sources to the UNC system.

References Cited

Baldwin, Christopher, and Penny Pasque. 2006. "The Benefits of College: Public and Private, Economic and Social." *Perspectives: Policy Edition*, September.

Crook, Clive. 2006. "A Matter of Degrees: Why College is Not an Economic Cure-All." *The Atlantic Monthly*, 298(4):25-30.

Holland, Dorothy and Margaret Eisenhart. 1990. *Educated in Romance: Women, Achievement, and College Culture*. Chicago: University of Chicago Press.

Nathan, Rebekah. 2005. *My Freshman Year: What a Professor Learned by Becoming a Student*. Ithaca, NY: Cornell University Press.

Wilson, Robin. 2006. "Study Links Proportion of Part-Time Instructors with Graduation Rates at 2-Year Colleges." *Chronicle of Higher Education* online daily news, 17 October.