Looking Where the Light Is Better: A Review of the Literature on Assessing Higher Education Quality[1]

By Marc Chun, fellow, RAND Corporations's Council for Aid to Education

An old joke recounts how a woman notices a man on his hands and knees while he frantically searches for something under a streetlamp. "Excuse me?" she asks. "Do you need some help?"

"Oh, yes, I'm looking for my car keys," he replies, and gestures towards his idle car in the darkness half a block away.

As she kneels down to assist, she inquires, "Where exactly did you lose the keys?"

As he carefully scans the pavement around him, he points off down the block and replies, "Over there by the car."

She pauses and shoots him a quizzical look. "Then why are you looking over here?" she queries.

"The light's better."

Given the most recent push for assessing higher education quality (framed in the public policy discourse as an issue of accountability), it is instructive to review the research literature, which demonstrates that there has been tremendous ongoing assessment effort in the United States over the past forty years. This assessment has occurred simultaneously at multiple levels. At the state level, recent research found that, by 1997, more than three-quarters of the states had some form of higher education assessment policy; however, the researchers note that "little systemic knowledge has been available to measure the extent and scope of publicly mandated outcomes assessments" (Nettles, Cole and Sharp 1997). At the institutional level, all institutions engaged in some form of assessment (often linked to self-studies for accreditation purposes); however, of the 1,393 public and private institutions recently surveyed, 82 percent listed "Excellence in Undergraduate Education" as part of their mission statement, but 38 percent did not conduct studies to link student experiences to student outcomes (Peterson, Augustine, Einarson and Vaughan 1999a, 1999b, 1999c).

Indeed, the literature shows that much has been (and can be) learned from both the state-level and institution-level assessment efforts. But if we take as our starting point that one of the central purposes of higher education is student learning, the obvious question arises: Are we indeed measuring what we *should* be measuring? Or, to what extent do we measure what is easier to measure? Are we looking merely where the light is better?

Four Approaches to Data Collection

The methodological approaches traditionally used to

[1] *This article uses material from a preliminary literature review completed by RAND Associate Researcher Catherine Augustine.*

assess higher education quality can be organized into four basic families or groupings: (1) actuarial data; (2) ratings of institutional quality; (3) student surveys; and (4) direct measures of student learning. Each will be discussed separately.

Actuarial Data

What are often seen as the most "objective" measures of higher education quality are the analyses based on "actuarial" data. These data include graduation rates, racial/ethnic composition of the student body, level of endowment, student/faculty ratio, highest degree earned by faculty members, breadth and depth of academic course offerings, selectivity ratio, admissions test scores of entering students, and levels of external research funding. Researchers argue that the primary advantages are that these data are relatively straightforward to collect, and the resulting statistics can be easily compared across institutions and over time. Although not intrinsic to the data themselves, the way in which the analyses are conducted typically relies upon a central assumption: A better quality educational institution (or a better quality educational experience) is necessarily associated with more and better resources—in this case, better funding, better faculty (which is defined as a higher percentage of any given cadre holding Ph.D.s), and better students (which is operationalized as resulting from higher admissions selectivity) (Astin 1968, 1977, 1991, 1993).

Actuarial data have been used by some states to measure institutional effectiveness (NCHEMS 1994). For example, the Texas Higher Education Coordinating Board gathers data in order to track students. As part of the ongoing review of two-year colleges, the coordinating board has developed the Academic Performance Indicator System (Gates et al. 2001). This information system contains longitudinal data on courses and students (demographic information, Social Security numbers, course enrollment, and graduation and Texas employment status), which allows students to be tracked across colleges and into the workforce by linking Social Security numbers to Texas workforce commission data.

Other examples of actuarial approaches include the National Center for Education Statistics and the Integrated Postsecondary Education Data System, which include data on student enrollment, faculty ranks, and institutional expenditures. These national databases are huge in scope, and some of the data come from secondary sources—such as census counts and transcripts (NCHEMS 1994). However, recent reviews of national data systems concluded that current databases yield little information about an institution's educational effectiveness in terms of the student outcomes it produces (Dey et al. 1997; NPEC 2000). In addition, a 1999 study found that only 10 percent of the approximately 1,300 institutions responding to a national survey reported having an institutional database that linked student information with faculty, curricular, and

References

Aakar, D. A., V. Kumar and G. S. Day. 1998. *Marketing research*. New York: Wiley.

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* 40:5, 499-526.

Astin, A. W. 1993. *What matters in college? Four critical years revisited*. San Francisco: Jossey-Bass.

——. 1992. *Cognitive development among college undergraduates*. Doctoral Dissertation. Los Angeles: University of California, Los Angeles.

——. 1991. *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. New York: American Council on Education/ Macmillan.

——. 1977. *Four critical years: Effects of college on beliefs, values, and knowledge*. San Francisco: Jossey-Bass.

——. 1968. Undergraduate achievement and institutional "excellence." *Science* 161, August, 661-668.

Baird, L. L. 1976. *Using self-reports to predict student performance*. New York: The College Board.

Banta, T. W., E. W. Lambert, G. R. Pike, J. L. Schmidhammer and J. A. Schneider. 1987. Estimated student score gain on the ACT Comp Exam: Valid tool for institutional assessment. *Research in Higher Education* 273, 195-217.

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.

Berdie, R. F. 1971. Self-claimed and tested knowledge. *Educational and psychological measurements* 31, 629-636.

Black, S. 1993. Portfolio Assessment. *The Executive Educator* 15, 28-31.

financial databases (Peterson et al. 1999).

The literature indicates, then, that in all of these cases, although actuarial data have prima facie validity in objectively assessing higher education quality, it is not clear if the analyses can even tacitly measure student learning.

Ratings of Institutional Quality

A second approach is based on analyses of ratings and rankings of institutional quality. This has typically taken the form of surveying either or both college faculty and administrators and asking these "experts" to rate the quality of different institutions and their programs on a series of dimensions. The implicit logic here is that informed "experts" can best assess institutional quality.

Perhaps the best-known (and most notorious) use of such analyses is the annual college rankings published by *U.S. News & World Report*, which have become the best-selling college guide in the United States. The rankings are based in part on actuarial data (such as selectivity, faculty resources, and financial resources), but are also based on surveys of faculty and administrators on their perceptions and opinions about academic quality and reputation. Although the general approach of using of multiple indicators and measures is consistent with the assessment literature (e.g., see Riggs and Worthley 1992; Astin 1991; Ewell 1984, 1988b; Gentemann et al. 1994; Halpern 1987; Jacobi et al. 1987; Ratcliff, Jones et al. 1997; Terenzini 1989; Vandament 1987), the *U.S. News & World Report* rankings have come under fire for a number of reasons.

Of primary concern have been the methods used to calculate the rankings. A 1977 report by the National Opinion Research Center (NORC)—commissioned by *U.S. News & World Report*—presented a systematic review of the methods used in the rankings. The NORC report notes that "the principal weaknesses of the current approach is that the weights used to combine the various measures into an overall rating lack any defensible empirical or theoretical basis. Recent studies of the measure by McGuire (1995) and Machung (1995) indicate that the ratings are sensitive to relatively small changes in the weighting scheme." The *U.S. News* weighting scheme is difficult to defend, and the NORC study concludes that, "since the method of combining the measures is critical to the eventual ratings, the weights are the most vulnerable part of the methodology." NORC also notes that a simple correlation matrix of the variables is not presented, which would indicate whether or not the measures are collinear and are, in essence, measuring the same thing. They also note that some variables may lack face validity. Alumni giving is claimed to serve as a proxy for satisfaction, when it can arguably be instead a function of effectiveness of the development office.

The NORC study also notes that reputational ratings play a huge role in rankings (college presidents are asked to rank other institutions), but it is questionable whether or not the respondents are able to make judgments about such a wide range of insti-

Bohr, L., E. Pascarella, A. Nora, B. Zusman, M. Jacobs, M. Desler and C. Bulakowski. 1994. Cognitive effects of two-year and four-year institutions: A preliminary study. *Community College Review* 22:1, 4–11.

Bradburn, N. M., and S. Sudman. 1988. *Polls and surveys: Understanding what they tell us*. San Francisco: Jossey-Bass.

Brandt, R. M. 1958. The accuracy of self estimates. *Genetic psychology monographs* 58, 55-99.

Cole, J. J. K., M. T. Nettles and S. Sharp. 1997. *Assessment of teaching and learning for improvement and accountability: State governing, coordinating board and regional accreditation association policies and practices*. Ann Arbor: University of Michigan, National Center for Postsecondary Improvement.

Converse, J. M. and S. Presser. 1989. *Survey questions: Handcrafting the standardized questionnaire*. Newbury Park, CA: Sage.

DeNisi, A. S. and J. B. Shaw. 1977. Investigation of the uses of self-reports of abilities. *Journal of Applied Pscyhology* 62, 641-644.

Dey, E., S. Hurtado, B. Rhee, K. K. Inkelas, L. A. Wimsatt, F. Guan. 1997. *Improving Research on Postsecondary Outcomes: A Review of the Strengths and Limitations of National Data Sources*. Stanford, CA: National Center for Postsecondary Improvement.

Ewell, P. T. 1988. Outcomes, assessment, and academic improvement: In search of usable knowledge. In J. C. Smart, ed. *Higher education: Handbook of theory and research* 4, 53-108. New York: Agathon Press.

——. 1987. Establishing a campus-based assessment program. In D. F. Halpern, ed. Student outcomes assessment: What institutions stand to gain. *New Directions for Higher Education* 59, 9-24.

——. 1984. *The self-regarding institution: Information for excellence*. Boulder, CO:

tutions. As noted in the study, "The large number of institutions within each classification means that each rater is asked to rate about 2000 institutions."

Moreover, the underlying assumptions about reputation may also be of concern. The NORC study notes, "The principle [sic] limitations are its inherently subjective nature and the fact that academic excellence, at least as traditionally defined, is not the goal of all, or perhaps even the majority, of colleges or students. In addition, it is generally assumed that reputations change more slowly than real change in institutions, thus overvaluing institutions that, in fact, may be declining and undervaluing institutions that are improving."

In addition, the *U.S. News* approach does not measure what many claim to be the most important measure of programmatic and institutional effectiveness: directly measured student abilities (Winter, McClelland and Stewart 1981; Graham and Thompson 2001). The NORC study concludes that, in addition to a need to measure student experiences, "the other area that is absent from the current set of measures relates to the academic demands of the curriculum…. There is not a good taxonomy of curricula, and the literature in this area is not particularly helpful." It should be noted that, in 1996, *U.S. News* added a measure of "value added," which they defined as the difference between actual and expected graduation rates. Such an operationalization is highly problematic. A high or low graduation rate may have drastically different meanings in different contexts, and the measure has no direct link to what students have actually learned at the institution.

In an article in the *Washington Monthly*, editor Nicolas Thomson writes, "A single magazine's idiosyncratic ranking system may seem peripheral to the larger issues of higher education, but this particular one matters a lot ... the rankings do have a kind of Heisenberg effect, changing the very things they measure and, in certain ways, changing the entire shape of higher education. The problem isn't that the rankings put schools in the wrong order ... a better ranking system ... would push [a school] to become an even better school…. Unfortunately, the *U.S. News* rankings instead push schools to improve in tangential ways and fuel the increasingly prominent view that colleges are merely places in which to earn credentials." Why do the rankings have such widespread acceptance? Thompson writes, "The rankings are opaque enough that no one outside the magazine can figure out exactly how they work, yet clear enough to imply legitimacy."

National Center for Higher Education Management Systems.

Ewell, P. T. and D. P. Jones. 1993. Actions matter: The case for indirect measures in assessing higher education's progress on the national education goals. *Journal of General Education* 42, 213-148.

Fong, B. 1988. Assessing the departmental major. In J. H. McMillan, ed. *Assessing students' learning*. New Directions for Teaching and Learning 34, 71-83. San Francisco: Jossey-Bass.

Forrest, A. and A study group on portfolio assessment. 1990. *Time will tell: Portfolio-assisted assessment of general education*. The AAHE Assessment Forum, American Association for Higher Education.

Gamson, Z. F. and S. J. Poulsen. 1989. Inventories of good practice: The next step for the seven principles for good practice in undergraduate education. *AAHE Bulletin* 42, 7-8.

Gates, S. M., C. H. Augustine, R. Benjamin, T. K. Bikson, E. Derghazarian, T. Kaganoff, D. G. Levy, J. S. Moini and R. W. Zimmer. 2001. *Ensuring the quality and productivity of education and professional development activities: A review of approaches and lessons for DoD*. Santa Monica, CA: National Defense Research Institute, RAND.

Gentemann, K. M., J. J. Fletcher and D. L. Potter 1994. Refocusing the academic program review on student learning. In M. K. Kinnick, ed. *Providing useful information for deans and department chairs*. New Directions for Institutional Research 84, 31-46. San Francisco: Jossey-Bass.

Gill, W. E. 1993. Conversations about accreditation: Middle States Association of Colleges and Schools: Focusing on outcomes assessment in the accreditation process. Paper presented at Double Feature Conference on Assessment and Continuous Quality Improvement of the American Association for Higher Education, Chicago, IL. ERIC

Indeed, it is questionable whether or not the rankings have changed educational practices at the institutional level. In a survey of nearly 1,400 colleges and universities, Peterson and Augustine (2000) attempted to determine if assessment was used as an end itself or if it was used to improve education. They concluded that "student assessment has only a marginal influence on academic decision making" and that faculty members involved in governance were supportive of assessment at only a quarter of these institutions. The greatest impact, then, seems to be how the rankings shift student application patterns. A study by Monk and Ehrenberg (1999) for the National Bureau of Economic Research found that moving up one place in an institution's ranking results in an increase in admittance rate of 0.4 percent.

As a result, many have rejected the meaningfulness of the rankings and their usefulness in shaping educational and curricular policy to improve student learning. According to Donald Kennedy, president of then-first-ranked Stanford, "It's a beauty contest, not a serious analysis of quality." In 1998, *The New York Times* reported that law schools mailed pamphlets titled "Law School Rankings May Be Hazardous to Your Health" to 93,000 law school applicants.

It is undeniable that institutional rankings have a widespread impact on the college-going behavior of student applicants, on institutional programmatic changes (in an attempt to move up in the rankings), and in reinforcing cultural assumptions about what constitutes a quality undergraduate experience. Again, however, the literature demonstrates that there is no clear link between such rankings and actual student learning.

Student Surveys

A third approach used to measure institutional quality is based on self-reported student information. In contrast to the proxy data used in the actuarial approach and ranking data based on surveying faculty and administrators, these data are collected by asking students directly about their collegiate experiences, satisfaction with their coursework and school, self-assessments of improvement in their academic abilities, and educational and employment plans.

The two most common methods for gathering such data are through surveys (Astin 1991; Ewell 1987c; Gill 1993; Johnson et al. 1993; Lenning 1988; Muffo and Bunda 1993) and interviews of individuals or groups (Johnson et al. 1993; Lenning 1988; Smith et al. 1993), which in some cases may supplement student interviews with those of faculty and other stakeholders. Ostensibly, the methodological advantage of these surveys is that data can economically be collected on a large-scale basis. Individual institutions collect such data to gather feedback about their institution (NCHEMS 1994), and national researchers collect data from a number of institutions in order to generate research on the effects of higher education in general and on the between-college impacts. Such self-reported information has also been used in an

Document Reproduction Service No. ED 358 792

Graham, A. and N. Thompson. 2001. Broken ranks: U.S. News' college rankings measure everything but what matters. And most universities do not seem to mind. *The Washington Monthly*, September. Available at: www.washingtonmonthly.com/features/2001/0109. graham.thompson.html.

Halpern, D. F. 1987. Recommendations and caveats. In D. F. Halpern, ed. *Student outcomes assessment: What institutions stand to gain*. New Directions for Higher Education 59, 109-111.

Hansford, B. C. and J. A. Hattie. 1982. The relationship between self and achievement/performance measures. *Review of Educational Research* 52, 123-142.

Hutchings, P. 1989. *Behind outcomes: Contexts and questions*. The AAHE Assessment Forum, American Association for Higher Education.

Jacobi, M., A. Astin and F. Ayala. 1987. *College student outcomes assessment: A talent development perspective*. ASHE-ERIC Higher Education Report No. 7. Washington, DC: Association for the Study of Higher Education.

Johnson, R., R. D. McCormick, J. S. Prus and J. S. Rogers. 1993. Assessment options for the college major. In T. W. Banta and Associates, eds. *Making a difference: Outcomes of a decade of assessment in higher education*, 151-167. San Francisco: Jossey-Bass.

Koretz, D., B. Stecher, S. Klein and D. McCaffrey. 1994. The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and practices* 13: 3, 5-16.

Kuh, G. D. 2001. Assessing what really matters to student learning. *Change* 33:3,10-17.

attempt to assess institutional effectiveness (Astin 1993; Pace 1990; Terenzini and Wright 1987).

For example, the Baccalaureate and Beyond Longitudinal Study, which is based on the National Postsecondary Student Aid Study, gathers information about education and work experiences after student completion of the bachelor's degree. The study, which surveys a nationally representative sample of institutions, students, and parents, includes cross-sectional data gathered one year after bachelor's degree completion. Also included are longitudinal data regarding entry into and progress through graduate level education and the workforce. The goal is to follow each cohort over a twelve-year period.

The National Survey of Student Engagement is an annual student survey designed to aid colleges and universities in improving student learning (Kuh 2001). The survey assesses the extent to which students from approximately 470 four-year colleges and universities participate in activities associated with learning and development. Kuh notes that a goal of the project is to change the way people think and talk about higher education quality.

The Cooperative Institutional Research Program (CIRP) survey, administered by UCLA's Higher Education Research Institute, is touted as the most comprehensive, longest, and largest higher education student survey. Annual data collection began in 1966, and the fall 2000 administration of the Freshman Survey included 717 participating institutions nationwide and over 404,000 students (which is almost a quarter of the nearly 1.64 million first-time, full-time first year students). The Follow-Up Survey is typically given to a sub-sample of students eight years after entering college. The surveys utilize self-reports on activities and goals as well as self-ratings. The assumption underlying self-reported data is that respondents can describe their feelings (such as satisfaction), their behaviors (such as time-on-task), and their opinions. In addition, it is assumed that students can describe their current abilities as well as their learning gains or improvements over time. Faculty members and administrators have also been surveyed about their feelings, behaviors, and opinions (Peterson 1987; Gamson and Poulsen 1989). Astin and his colleagues (1991) developed a survey of faculty (UCLA's Higher Education Research Institute Faculty Survey) that includes items on teaching techniques and assessment methods. These self-reports have been used

> Although it may seem to be the most obvious way to assess the quality of undergraduate education, the use of direct measures of student learning is uncommon.

Laing, J., R. Swayer, and J. Noble. 1989. Accuracy of self-reported activities and accomplishments college-bound seniors. *Journal of College Student Development* 29, 362-368.

Lenning, O. T. 1988. Use of noncognitive measures in assessment. In T. W. Banta, ed. *Implementing outcomes assessment: Promise and perils*. New Directions for Institutional Research 59, 41-51. San Francisco: Jossey-Bass.

Lowman, R. L., and R. E. Williams. 1987. Validity of self-ratings of abilities and competencies. *Journal of Vocational Behavior* 31, 1-13.

Machung, A. 1995. Changes in college rankings: How real are they? Paper presented at the 35th Annual AIR Forum, Boston, MA.

McGuire, M. D. 1995. Validity issues for reputational studies. In Walleri, R. D. and M. K. Moss, eds. *Evaluating and responding to college guidebooks and rankings*. New directions for institutional research 88, Winter. San Francisco: Jossey-Bass.

Monk, J. and R. G. Ehrenberg. 1999. U.S. News and World Report rankings: Why do they matter. *Change*, November/December, 43-51.

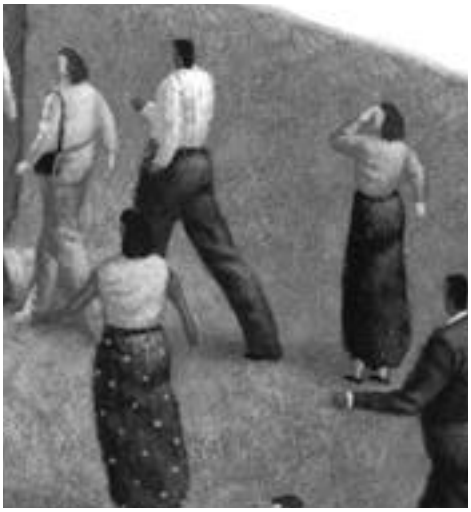
Muffo, J. A. and M. A. Bunda. 1993. Attitude and Opinion Data. In Banta, Trudy and Associates, eds. *Making a difference: Outcomes of a decade of assessment in higher education*, 139-150. San Francisco: Jossey-Bass Publishers

National Center for Higher Education Management Systems. 1996. *The national assessment of college student learning: An inventory of state-level assessment activities*. Boulder, CO: National Center for Higher Education Management Systems.

—. 1994. *A preliminary study of the feasibility and utility for national policy of instructional and good practice indicators in undergraduate education*. Contractor Report for the National Center for Education Statistics. Boulder, CO: National

in student assessment efforts (Pascarella and Terenzini 1991).

A key issue in student surveys, as in all surveys, is that of the reliability of the self-reported data. Because many of the outcomes of interest cannot be empirically measured (e.g., attitudes and values), the use of student self-reports is commonplace in higher education research. Researchers have studied the credibility of these self-reports (Berdie 1971; Pohlman and Beggs 1974; Baird 1976; Turner and Martin 1984; Pace 1985; Pike 1995; and Ouiment, et al. 2001) and, as noted by Kuh (2001), there are two problems that impact the accuracy of self-reports. First, some respondents are *unable* to supply accurate information; and second, some respondents are *unwilling* to supply accurate information (Wentland and Smith, 1993; Aaker, Kumar and Day 1998). Either condition clearly affects the efficacy of the data and the subsequent analyses. However, Pike (1999) also studied the "halo effect," in which student respondents may inflate reporting of their behavior, performance, or what they perceive they have gained from their college experience towards the more socially acceptable. He argues that, because the effect is consistent across students and institutions, comparisons are not compromised. (This, however, is still a concern when it comes to having an "accurate" picture of student growth.)



Again, one challenge in student surveys is ascertaining whether or not what students report corresponds to what they actually experienced. Ouiment et al. (2001) considered student responses to the College Student Report. They used focus groups and survey instruments together and concluded that, although there was some variation in respondents' interpretation of some items on the survey, there was a general consensus for a "vast majority of items." They also concluded that "the meaning of the response categories were item specific; that is, the meaning of 'very often' to one question did not necessarily represent the same frequency as another item."

However, other research suggests that student surveys may nonetheless be a viable approach. Some researchers found that self-reports are highly correlated with quantifiable measures of student progress (Anaya 1992; Anaya 1999; Dumont and Troelstrup 1980; Ewell and Jones 1993). Furthermore, Astin's (1993) studies on the relationship between self-reported data and student achievement indicate that the patterns of self-reported data vary by major and student experiences in ways that mirror the patterns found by directly assessing cognitive outcomes.

Center for Higher Education Management Systems.

National Opinion Research Center. 1987. *A review of the methodology for the* U.S. News & World Report*'s rankings of undergraduate colleges and universities*. Report by the National Opinion Research Center.

National Postsecondary Education Cooperative. 2000. *The NPEC sourcebook on assessment, volume 1: Definitions and assessment methods for critical thinking, problem solving, and writing*. Center for Assessment and Research Studies, James Madison University, Harrisonburg, VA, under the sponsorship of the National Center for Education Statistics, U.S. Department of Education.

——. 2000. *The NPEC sourcebook on assessment, volume 2: Selected institutions utilizing assessment results*. Center for Assessment and Research Studies, James Madison University, Harrisonburg, VA, under the sponsorship of the National Center for Education Statistics, U.S. Department of Education.

Nettles, M. T., J. J. K. Cole and S. Sharp. 1997. *Assessment of teaching and learning in higher education and public accountability*. Stanford, CA: National Center for Postsecondary Improvement.

Obler, S. S., J. Slark and L Umbdenstock 1993. Classroom assessment. In T. W. Banta and Associates, eds. *Making a difference: Outcomes of a decade of assessment in higher education*, 211-226. San Francisco: Jossey-Bass.

Ouiment, J. A., R. M. Carini, G. D. Kuh, and J. C. Bunnage. 2001. Using Focus Groups to Establish the Validity and Reliability of a College Student Survey. Paper presented at 2001 AIR Forum, Long Beach, CA.

Pace, C. R. 1990. *The undergraduates; A report of their activities and progress in college in the 1980s*. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.

Kuh (2001) concludes, based on his review of the research (Bradburn and Sudman 1988; Brandt 1958; Converse and Presser 1989; DeNisi and Shaw 1977; Hansford and Hattie 1982; Laing, Swayer and Noble 1989; Lowman and Williams 1987; Pace 1985; Pike 1995), that self-reports are valid under five conditions: "(1) when the information requested is known to the respondents; (2) the questions are phrased clearly and unambiguously; (3) the questions refer to recent activities; (4) the respondents think the questions merit a serious and thoughtful response; and (5) answering the questions does not threaten, embarrass, or violate the privacy of the respondents or encourage the respondent to respond in socially desirable ways."

Setting aside the difficulties in data collection, another concern has been raised about analysis of student survey data. Often, as in analyses of the CIRP data, researchers rely on a central conceptual paradigm that one can assess the impact of college using essentially the pre- and post-test model. Although having two time points clearly has advantages over a solely retrospective survey design, it is nonetheless problematic to determine the actual impact of any process variables. Moreover, the traditional positivistic approach often employed in such analyses assumes that individual aspects of the college experience can be studied atomistically; it can be seen as denying the holistic nature of the college experience.

Thus, although student surveys can and have been used in an attempt to link educational quality with student learning, their use is problematic specifically in assessing student learning because of the indirect measure of learning given the reliance on student self-assessment.

Direct Assessments of Student Learning

A fourth approach to assess institutional quality is to measure student learning directly. Direct assessments of student learning are perhaps the least systematically used of the four approaches discussed here. This may involve analyzing course grades; administering standardized tests, performance tasks, and special multiple-choice or open-ended tests to assess general academic skills or subject matter knowledge; and obtaining data from other measures, such as evaluations of student projects, portfolios of student work, etc.

Some researchers have used direct measures of student learning as a means of collecting data on programmatic and institutional effectiveness (Winter, McClelland and Stewart 1981). However, most of these efforts are conducted by an institution's faculty and staff on their own students. As a result, comparisons between institutions are less common (exceptions include Bohr et al. 1994; Pascarella et al. 1994). Still, some institutions have collaborated in directly measuring student learning outcomes in order to compare results among themselves (Obler et al. 1993). In addition, some states have required that all institutions use the same standardized measures in directly assessing students' knowledge, skills, and

——. 1985. *The credibility of student self-reports*. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.

Palomba, C. A. and T. W. Banta. 1999. *Assessment essentials: Planning, implementing, and improving assessment in higher education*. San Francisco: Jossey-Bass.

Pascarella, E. T., L. Bohr, A. Nora and P.T. Terenzini. 1994. *Is differential exposure to college linked to the development of critical thinking?* Illinois Univ., Chicago: National Center on Postsecondary Teaching, Learning, and Assessment.

Pascarella, E. T., and Terenzini, P. T. 1991. *How college affects students: Findings and insights from twenty years of research*. San Francisco: Jossey-Bass.

Peterson, M. W. 1987. *Academic management practices survey for the research program on the organizational context for teaching and learning*. Ann Arbor: National Center for Research to Improve Postsecondary Teaching and Learning, The University of Michigan.

Peterson, M. W. and C. H. Augustine. 2000. The influences of regional accreditation associations on institutions' approaches to, support for, and use of student assessment. In *A collection of papers on self-study and institutional improvement*. Prepared for the program of the Commission on Institutions of Higher Education at the 105th Annual Meetings of the North Central Association of Colleges and Schools, April 1-4, 2000, Chicago.

Peterson, M., C. H. Augustine, M. K. Einarson, D. S. Vaughan. 1999a. *Designing student assessment to strengthen institutional performance in associate of arts institutions*. Stanford, CA: National Center for Postsecondary Improvement.

——. 1999b. *Designing student assessment to strengthen institutional performance in comprehensive institutions*. Stanford, CA:

abilities (Cole et al. 1997; NCHEMS 1996; Steele and Lutz 1995). These methods have been used to collect data on individual students and on groups of students at both the program and institutional levels (Ratcliff, Jones, Guthrie and Oehler 1991).

In addition to the more standard and commonly used paper and pencil examinations, direct assessments of students can also be done through evaluating on-demand student performances, such as presentations, debates, dances, and musical recitals (Palomba and Banta 1999). These performances can be evaluated at the end of a student's career in order to assess programmatic effectiveness. Researchers tend to agree on the validity of this approach in terms of measuring students' abilities, but the use of one performance may not be reliable. For example, a student may write an excellent term paper on one topic, but not on another, due to varying levels of motivation or interest in the topic. Such a lack of consistency may not be important, however, if the goal is to evaluate the effectiveness of the program rather than of the student (Johnson et al. 1993; Lenning 1988). The evaluation process tends to be low-cost to the institution, although students may expend a great deal of resources in completing the long-term projects. While students may enter their projects in state or national competitions, there is little evidence that these projects are compared in order to make judgments about program effectiveness across institutions. Such comparisons could be difficult due to variations in curriculums between institutions.

While the importance and value of student learning are generally accepted, few agree on how best to assess it.

In order to overcome the problem of reliability with some of these direct measures, scholars have advocated the use of portfolios (Banta et al. 1996; Black 1993; Forrest 1990; Hutchings 1989; Suen and Parkes 1996). Portfolios require students to assemble cumulative samples of their work products and often include a self-evaluative component (Black 1993; Fong 1988; Johnson et al. 1993; Waluconis 1993). While evaluating multiple student products overcomes problems of reliability, validity concerns remain. It is difficult to ensure that the work presented in a portfolio represents only the work of the student. If results of group work are allowed in the portfolio, it is again difficult to ascribe the work to the student. Moreover, Koretz et al. (1994) argue that portfolio assessments are unreliable.

Still, some have argued that direct assessment can be used as a means of academic accountability and as a tool for curriculum reform and institutional evaluation (Mingle 1986). For example, the Texas

National Center for Postsecondary Improvement.

——. 1999c. Designing student assessment to strengthen institutional performance in doctoral and research institutions. Stanford, CA: National Center for Postsecondary Improvement.

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

——. 1995. The relationship between self reports of college experiences and achievement test scores. *Research in Higher Education* 36, 1-22.

Pohlman, J. T., and D. L. Beggs. 1974. A study of the validity of self-reported measures of academic growth. *Journal of Educational Measurement* 11, 115-119.

Ratcliff, J. L., E. A. Jones, et al. 1997. *Turning results into improvement strategies*. University Park: The Pennsylvania State University, National Center on Postsecondary Teaching, Learning, and Assessment.

Ratcliff, J. L., E. A. Jones, D. S. Guthrie and D. Oehler. 1991. *The effect of coursework patterns, advisement, and course selection on the development of general learned abilities of college graduates*. University Park: The Pennsylvania State University, National Center on Postsecondary Teaching, Learning, and Assessment.

Riggs, M. L. and J. S. Worthley, Baseline Characteristics of Successful Program of Student Outcomes Assessment. ERIC document ED353285

Smith, M. K., J. L. Bradley and G. F. Draper. 1993. *A national survey on assessment practices*. Knoxville, TN: University of Tennessee, Knoxville, Clearinghouse for Higher Education Assessment Instruments.

Steele, J. M. and D. A. Lutz 1995. *Report of ACT's research on postsecondary assessment needs*. Iowa City, IA: American College Testing Program.

Academic Skills Program is administered to all first-time freshmen and to all rising juniors as a means to ensure that all students attending public institutions of higher education have the basic skills for college-level work.

Although it may seem to be the most obvious way to assess the quality of undergraduate education, the use of direct measures of student learning is uncommon. The literature suggests several reasons for this. These approaches can be cost-prohibitive to implement, for example. And there are huge obstacles to making institutional comparisons. The most insurmountable of these is the need for institutions to agree on what should be measured.

Is There Madness to the Methods?

When it comes to understanding what student have actually learned in college (and linking learning to assessments of institutional quality), the literature suggests that we are faced with a conundrum. While the importance and value of student learning are generally accepted, few agree on how best to assess it. The literature further suggests that this can be better understood by considering the available methods.

Actuarial data is commonly used because of the ease of collection and the patina of scientific objectivity. But this approach equates quality with discrete, available, and, perhaps most significantly, easily *measurable* indicators of quality, such as counts of people and resources. Institutional rankings rely on a formula that combines actuarial data and ratings by informed experts. These rankings are limited (and questionable) because they provide only an indirect measure of quality and conflate quality and reputation. Student surveys have attempted to measure quality using student perceptions of their learning. Research has shown, however, that such measures may be problematic because they depend upon student self-evaluation. Still, this research has been an important step in connecting student learning with educational quality. And finally, while direct measures of student learning may arguably have the greatest face validity with regard to assessing undergraduate education, the literature indicates that there are numerous implementation issues.

This last point is perhaps the most significant in a profoundly important yet subtle way. Whereas the discussions in the literature about the first three methods have debated whether or not these approaches *can* measure student learning (and question whether or not the proxies used are valid or appropriate), discussions about the direct measures of student learning debate *how* student learning should best be done. Granted, these debates are perhaps just as fierce: At what point should students be assessed? What should be included in the assessment? What is the best means to collect the information? And how can it be ensured that these data are reliable? The central point, however, is that few would deny that direct measures of learning are an appropriate means to assess the quality of undergraduate education. In other words, if we are interested in understanding what students have learned, we should measure what students have learned. The key is to focus on developing better methods to directly assess student learning. Thus, to return to the anecdote that opened this discussion, we know where we should be looking. We will find the keys by building a better streetlight.

Suen, H. K. and J. Parkes. 1996. *Challenges and opportunities for student assessment in distance education*. Distance Education Online Symposium, 6 7 [On-line serial]. Available: Internet: ACSDE@ PSUVM.PSU.EDU.

Terenzini, P. T. 1989. Assessment with open eyes: Pitfalls in studying student outcomes. *Journal of Higher Education* 60, 644-664.

Terenzini, P. T. and T. Wright. 1987. Influences on students' academic growth during four years of college. *Research in Higher Education* 26, 161-179.

Thomson, Nicolas. 2000. Playing with numbers: How U.S. news mismeasures higher education and what we can do about it. *Washington Monthly,* September.

Turner, C. F. and E. Martin, eds. 1984. *Surveying subjective phenomena, Vol 1.* New York: Russell Sage Foundation.

Vandament, W. E. 1987. A state university perspective on student outcomes assessment. In D. F. Halpern, ed. *Student outcomes assessment: What institutions stand to gain*. New Directions for Higher Education 59, 25-28.

Waluconis, Carl J. 1993. Student self-evaluation. In Trudy Banta, ed. *Making a difference: Outcomes of a decade of assessment in higher education*, 244-255. San Francisco: Jossey-Bass.

Wentland, E. J. and K. W. Smith. 1993 *Survey responses: An evaluation of their validity*. New York: Academic Press.

Winter, D. G., D. C. McClelland and A. J. Stewart. 1981. *A new case for the liberal arts*. San Francisco: Jossey-Bass.