100 mm
A5
1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz
1234567890
1.0 mm
1.5 mm
2.0 mm

DOCUMENT RESUME

ED 284 889 TM 870 436

TITLE	Assessing the Outcomes of Higher Education. Proceedings of the ETS Invitational Conference (47th, New York, New York, October 25, 1986).
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ISBN-0-88685-062-2
PUB DATE	87
NOTE	114p.; For individual papers see TM 870 437-444 and ED 281 400.
AVAILABLE FROM	Invitational Conference Proceedings, Educational Testing Service, Princeton, NJ 08541-0001.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Academic Standards; Accountability; Accreditation (Institutions); College Students; *Educational Assessment; *Educational Objectives; Educational Policy; *Educational Testing; Government Role; *Higher Education; *Institutional Evaluation; *Measurement Objectives; *Outcomes of Education; Testing Problems; Testing Programs; Test Use; Test Validity
IDENTIFIERS	Kuder (Frederic); Value Added

ABSTRACT

Nine papers were presented at the 1986 Educational Testing Service Invitational Conference on outcomes assessment in higher education. The Award for Distinguished Service to Measurement was awarded to Frederic Kuder for development of the Kuder-Richardson 20 and KR-21 formulas for test reliability, the Kuder Preference Record, and the Kuder Occupational Interest Survey. The papers included a discussion of the goals and realities of American higher education (W. Ann Reynolds); the college's perspective of assessment (John W. Chandler); the state's perspective (Eleanor M. McMahon); and the accrediting association's perspective (Thurston E. Manning). Discussions also focused on critical validity issues in college assessment (Eva L. Baker); the case for unobtrusive measures (Patrick T. Terenzini); use of assessment to improve instruction (K. Patricia Cross); and value-added student assessment (Ernest T. Pascarella). In addition, Russell Edgerton presented a critical history of assessment in the form of a play. (GDC)

Assessing the Outcomes of Higher Education

Educational Testing Service

Assessing the Outcomes of Higher Education

Proceedings of the 1986 ETS Invitational Conference

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

The forty-seventh ETS Invitational Conference, sponsored by Educational Testing Service, was held at The Plaza, New York City, on October 25, 1986.

Presiding: Gregory R. Anrig
President
Educational Testing Service

Conference Coordinator: Margaret B. Lamb

Library of Congress Catalog Number: ISBN 0-88685-062-2

Contents

Introduction

With somewhat dramatic suddenness, the assessment of learning in college has emerged as an institutional, state, and national issue. It has been given particular urgency by legislation in several states requiring evidence of student progress and of the standards being met on individual campuses. Many colleges and universities are attempting to define objectives and identify appropriate ways to measure academic growth as well as attainment. The challenge to educational measurement is clear and immediate.

Old methodologies, often used by states to assess other levels of education, are not necessarily adequate to the complex task of assessing progress in higher education. The kinds of tests used to assess the basic skills of high school students, for instance, have little relevance to higher education, where the mastery of basic skills should be a prerequisite rather than a goal. W. Ann Reynolds warns, in her paper, against precipitous overreaction by state education officials. Moreover, as Eleanor M. McMahon suggests, statewide assessment efforts must be designed to preserve the autonomy and variety of our institutions of higher education.

During the next few years, therefore, those in higher education will find themselves under increasing pressure to develop new forms of assessment that reflect the advanced learning that takes place on college campuses. In the process, educators will have to define the goals and objectives of higher education in the late 20th century. Many of these goals will prove intangible and difficult to assess, and definitions may vary from campus to campus. Thurston E. Manning outlines the complexities of higher education assessment and calls for the development of tools that will provide a more accurate evaluation of institutional quality.

Some colleges, as indicated in many of the Invitational Conference papers, have already developed innovative programs that integrate assessment with education. These vanguard institutions include Northeast Missouri State University, which has made value-added testing an integral part of its program, and Alverno College, which has developed its own unique assessment program for defining and setting student educational objectives and monitoring their attainment. Ernest T. Pascarella suggests that the value-added approach can be helpful if it is applied with rigor.

A myriad of alternatives, many still undefined, remain to be explored. The directions assessment might take are suggested in papers by Eva L. Baker, who recommends that assessment focus on the individual educational experiences of students, and by Patrick T. Terenzini, who proposes an ingenious system of unobtrusive measures that would take advantage of existing yardsticks to gauge educational quality.

John W. Chandler suggests that academic departments might want to enlist the aid of unbiased outsiders in assessing student learning. K. Patricia Cross points out that, for assessment to have maximum impact, it should focus on the individual classroom.

As new assessment methodologies are developed, they must reflect any shifts that occur in the essential thrust of higher education. Russell Edgerton suggests that assessment methods geared to future societal demands should evolve from a view of higher education based not on a broad liberal arts curriculum but rather on the development of essential abilities, attitudes, and social skills, such as communicating effectively.

Collectively, the papers presented at this year's Invitational Conference, in the diversity of their recommendations, suggest that, despite the difficulty of devising a meaningful assessment of learning or educational quality in higher education, leaders in the field are already proposing imaginative and promising solutions. In the years ahead, ETS intends to work with concerned educators throughout the country in an effort to develop tools for higher education assessment that will meet the demands of the 21st century.

Gregory R. Anrig, President
Educational Testing Service

Presentation of the 1986 ETS Award for Distinguished Service to Measurement Citation

Presented to:
FREDERIC KUDER

Early in his long and distinguished career, just about 50 years ago, Frederic Kuder (in collaboration with Marion Richardson) derived formulas for test reliability in terms of item variances and covariances. This made it possible to estimate test reliability from item consistencies in a single test administration. Thus, in addition to alternate-form estimates of score equivalence and test-rated estimates of stability, test reliability could now be estimated in terms of internal item consistency without relying on arbitrary split-half methods. The resulting formulas, KR-20 for dichotomously-scored items and KR-21 for items roughly equal in difficulty, have dominated applied testing practice ever since. One consequence of this seminal contribution is that Dr. Kuder is the unwitting perpetrator of what has proven to be the most popular indoor sport in psychometrics—namely, rederiving the famous formula 20 from different or weaker assumptions.

Another interest that took root early in Dr. Kuder's career, and with which he has been occupied ever since, is the measurement of personal and occupational interests. In developing the *Kuder Preference Record* and its successor, the *Kuder Occupational Interest Survey*, he introduced a number of practical and theoretical innovations. These include the use of staggered pages to facilitate self-scorable answer sheets and, more fundamentally, the scoring of vocational interests directly in terms of similarity to occupational groups rather than indirectly in terms of differences between occupational groups and a general reference population. Through this latter insight, Dr. Kuder provided a firm psychometric foundation for a perennial problem in vocational counseling—not matching people to jobs in terms of predicted job performance but, rather, matching people to people in jobs in terms of anticipated job satisfaction.

For his landmark contributions to reliability theory and interest measurement, as well as for his decades of dedicated leadership as editor of both *Educational and Psychological Measurement* and *Personnel Psychology*, ETS is pleased to present its 1986 Award for Distinguished Service to Measurement to Frederic Kuder.

ETS Award for Distinguished Service to Measurement Recipients 1970-1986

1970 E. F. Lindquist

1971 Lee J. Cronbach

1972 Robert L. Thorndike

1973 Oscar K. Buros

1974 J. P. Guilford

1975 Harold Gulliksen

1976 Ralph W. Tyler

1977 Anne Anastasi

1978 John C. Flanagan

1979 Robert L. Ebel

1980 John B. Carroll

1981 Ledyard R Tucker

1982 Raymond B. Cattell

1983 Frederic M. Lord

1984 Louis Guttman
Henry Chauncey (special award)

1985 Paul Horst

1986 Frederic Kuder

Higher Learning in America: Aims and Realities

W. Ann Reynolds
Chancellor, The California State University

My undergraduate degree is a B.S. in Education, with majors in Biology and Chemistry and a minor in Theatre from Kansas State Teachers College, Emporia. The most *useful* (not the most inspiring nor challenging nor fun) course I took in college was tests and measurement, and later, in graduate school at Iowa, the same was true for biostatistics. Then, from 1974 through 1982, I served on the biology group for the Graduate Record Exam. They are meager but cherished credentials in testing which have served me well. Could it be that testing has the same aesthetics as Shaker furniture in that its beauty lies in its simple functionality? Please know I still yearn nostalgically for that isolated period of time I spent at Princeton each year with Bill Kastrinos and Gertrude Sanders and colleagues in biology, diligently slugging our way through new and old questions. I also miss the Canada geese on the lawn and the superb pie. The only thing I do not miss is the highway between the Newark airport and Princeton when it iced.

We have entered an era when many see testing as the answer to every problem—from AIDS to drug abuse to poor writing skills. Many of our state political leaders, and indeed the general public, place complete faith in testing. As a practicing test maker and teacher, I believe there are both considerable strengths and some limitations to the "quick fix" implied by testing.

No group is more aware of the *limits* of testing than you who are professional psychometrists and deal with testing issues on a daily basis. While in many ways the testing mania opens opportunities to develop further our expertise in the subtle, complex tasks of measuring human skills and knowledge, I urge that we be honest about the constraints on our abilities. This is especially true in measuring the learning that occurs in colleges and universities.

Before I pursue this theme further, I want to remind you how accurately the realities of higher education today match the aims of its early leaders

and sponsors. The nation has just helped to celebrate the 350th anniversary of its first university, Harvard. Its founders and those of other pioneer colleges, urged by religious leaders, established colleges in which ministers, teachers, and lawyers could be educated. In the next century this concern with education for the professions spread to provide an education for persons in technical and applied fields. In 1862, the federal government awarded land grants to states, providing them the funds to establish universities to teach agricultural and mechanical knowledge. Since then, states have increasingly used their own resources to provide education at public expense to their residents.

As the need for better-educated teachers in the early twentieth century became more apparent, states established colleges for teachers. The mission of all of these colleges has now been expanded so that a more comprehensive curriculum is offered, preparing students in a wider range of fields. As a result of this expansion the number of persons annually receiving a college education has expanded from the 12 in Harvard's first graduating class to about 1 million today in over 2,000 four-year colleges and universities. Ours is the first nation in the world to aspire to entitlement to a college degree essentially for all young people who qualify and desire to pursue the goal.

Recent studies comparing the United States' educational record with that of other countries have reflected ill on the United States with respect to the educational excellence of our schools. In contrast, the picture of the level of participation in higher education is actually quite bright. In 1980-81 some 32 percent of Americans 25 years and over had attained at least some college education. The closest international competitors were Canada and East Germany with 17 percent, and Japan and Sweden with 15 percent, respectively. By our estimate, today some 23.8 million people in the United States have a baccalaureate degree.

We have also tried with vigor over the last 25 years to increase the participation of women and minority students. The number of women and minorities attending college increased 85 percent between 1970 and 1983. In 1970, 3.5 million women were in college; in 1983 there were nearly 6.5 million. The number of persons attending part-time increased 88 percent and the number of students 25 years and older increased from 2.4 million to 5.1 million.

Happily for all of us in "the life," the projected decline in higher education enrollment during the 1980s simply has not materialized. Ten years ago, national conferences were devoted to planning for survival despite steady decreases in college enrollment. In the California State

University (CSU), with a 2.2 percent enrollment increase last year (8,600) and another 2 percent this year, we are scratching our heads about how to get more classrooms. Clearly we are witnessing an increase in the college-going rate.

Over the last two decades, public policy has demanded that, like elementary and secondary education, higher education be made available to persons regardless of their ability to pay. In 1963-64, financial aid to students in postsecondary education totalled about half a billion dollars. By 1981-82, it was more than $18 billion, about 80 percent from federal funds. The growth of student aid ended in the early 1980s, and now we face growing student indebtedness and an erosion of our ability to provide educational access to all who qualify.

Having met quantitative goals in providing more access and participation, higher education is now being asked to show that it has also met qualitative goals. In his impressions of democracy early in American history, Alexis de Toqueville asked if the leveling tendencies of broad participation in such social institutions as education and public life might not threaten their quality. Let us examine the relation of quantity and quality.

First, higher education has always been, in the eyes of many, the province of the elite—principally those who by the gift of social and economic status were aware of its value and could afford to pay for it—and of the intellectually elite—those who by virtue of native intelligence and a traditional educational experience could demonstrate eligibility for it. As opportunity is extended to large numbers of our population, many now say it cannot *really* be higher education. Too many seem convinced that improved access can only mean lower standards for higher education.

Second, the public and its political leaders are deeply concerned about our nation's economy and our increasing loss in competitive advantage. Many reports, making invidious comparisons between the educational performance of students and the vigor of the economy in the U.S. and, for instance, Japan, point to an inferior American educational system as the cause. And in an attempt to reverse this economic down spiral, the public appears once again to be willing to invest more heavily in higher education. But they count on that investment to yield a quickly visible economic return and regaining of total dominance in world trade—a heavy responsibility for higher education alone to achieve!

Third, the increasing use of accountability measures in kindergarten through twelfth grade has led the public and policy makers to think that

extension of use of such measures to college and university education might serve a useful purpose. The public indeed seems to fear that the "rising tide of mediocrity" that was said to threaten the public schools may also be lapping at the shores of our university campuses.

Let's take a moment to consider the concerns about accountability from the perspective of the persons most directly involved. You might call this the view from the lifeboat.

Faculty — Just as in real estate where the three most important considerations are location, location, location, the three most important issues for faculty, justifiably so, are salary, salary, salary. Running a close and urgent second is the issue of faculty development. Faculty need more time for scholarly pursuits, for travel (an ever-popular challenge to defend), for course planning and the honing of teaching and research skills.

Most faculty are truly eager to find ways to facilitate more learning in their classrooms. At the same time, they are mindful of academic freedom and their historical academic role in determining curricula and course content. Sometimes it is the little things. When I taught full-time, I was severely tried by the brimming chalk trays, containing pieces of chalk the size of a transistor. I still have my own personal box of colored chalk sequestered in my desk drawer as a talisman. And then there's parking. . . . .

Students — Students rightfully expect access to courses needed for a degree in the semesters or quarters that fit their plans. Our campuses are too profligate with student time as we line them up at bookstores, to see counselors, for degree checks. Their personal indebtedness has grown alarmingly. Nationwide, student borrowing increased from $2.3 billion in 1977-78 to an estimated $10.5 billion in 1985-86. It is not unusual for a CSU student to graduate with a $10,000 debt. (According to a recent survey of financial aid offices, the average indebtedness for students enrolled at a four-year public institution at the end of four years is $6,685. And at private institutions, the figure is $8,950. This rising indebtedness may also have channeled more college students into business or engineering careers than perhaps truly want to be there. In the CSU, for example, 16,500 undergraduate degrees (or 37 percent of the bachelor's degrees awarded) were conferred upon students who majored in these two popular areas. And then there's parking. . . . .

Trustees — Trustees feel inundated with information in contrast to the concern they might miss something that's going on for which they are

responsible. (At this very moment in the CSU with 19,000 faculty, 335,000 students and 9,000 support staff, I can assure you that somewhere somebody is doing something that is illegal, injudicious, harmful to well-being, or even all three of the above at once.) The vast majority of trustees yearn lovingly over the institutions they oversee. They come to agonize over the balances achieved between building and renovation needs, the salary benefit package for employees and liability insurance. Caring about an institution as they do, it is hard for them to see how fiduciary and policy responsibilities are translated into real growth in institutional quality plus student achievement.

Legislatures and Governors—In recent years education has been "in," but in a complicated, patchwork way. Nearly a hundred major studies (it is now "in" for educators to paper their guest powder rooms in draft copies of "A Nation at Risk" or "Who Will Teach Our Children?" rather than with wine bottle labels) have evolved, targeting one or more parts of the educational enterprise for reform. There is more coherence in their ontogeny than their output. Nearly every study starts with an organization that provides the topic or focus and finds fiscal support, usually from a private foundation or from federal or state government. A commission is appointed, chaired by the most famous person who will agree to take the job, and serving on it are a governor, a state or federal legislator, a college president, a dean, a faculty member, an emeritus distinguished educator, and then, always, Al Shanker. (I've now served on three national commissions with Al Shanker, attended most of the sessions, the reports are completed, and I've yet to meet him!) The target of all these studies has been policy makers, and usually those who control educational support dollars, hence mostly state legislators and governors.

The net effect has been spotty. Some 30 states now have implemented a stronger high school curriculum either as a graduation requirement or as a prerequisite for college, but much of this process antedates the reports. SAT and ACT scores have reversed their downward slide, banked, and show signs of slowly starting to climb again.

The urge to form a commission has found fertile ground in the subject of credentialing of teachers. We've just had a good example of the good and bad effects such a commission can have in California.

With half a million dollars from a private foundation, sponsorship by the chairs of two legislative committees on education and State Superintendent Bill Honig, as well as the blessing of other higher education leaders, a California Commission on the Teaching Profession was formed

over two and one-half years ago. The Commission labored long and hard with excellent staff support from Stanford University. Last December they produced a report proposing sweeping reforms to improve the teaching profession. The 27 recommended reforms had a price tag of some $1.7 billion. The report was lauded by organizations representing teachers, administrators, school board members, and legislators.

Then, the legislative session in January began with two major bills introduced to implement many of the 27 recommendations. These included reducing class size, lengthening the time to tenure for teachers, requiring publication of indices of conditions of teaching and learning in each school in the state, and a teacher credentialing process that introduced testing for beginning teachers and a critically important residency year.

Both bills began to undergo extensive hearings and amendments. Lobbyists hovered as in the landing pattern at La Guardia on a Friday evening. I'll bet you can guess what was the only pithy item left in that bill. You're right—testing for beginning teachers and they were to pay for it.

All the other far-reaching provisions for improving the teaching profession and making schools more accountable finally sifted down to establishing more hurdles for the yet unborn—prospective teachers. They were the only ones without an organization formed to oppose the reform bill. All the effort to think comprehensively about the problems of the teaching profession was lost in a last minute drive to get *something* passed that looked like educational reform.

A reform package of $1.7 billion had become a zero cost item—basically because the state had no money left for major reform. Genuine reform and sound recommendations were sacrificed to reality. The only voices in opposition to this focus on prospective teachers as a substitute for comprehensive reform of a vast network of professionals in complex environments with a clear need for additional state funding were those from higher education.

Interestingly, the good common sense of legislators ultimately prevailed. They saw that what had begun with the very best of intentions and the highest of motives had been reduced to a quick-fix solution that didn't gore anyone's ox—except fledgling teachers in a state that needs 90,000 of them in the next decade.

We now face the same challenge during this next legislative session. I am heartened that there is evidence of committment to comprehensive reform—at least before the session begins. I offer you the same opportu-

nity as those in the audience of *The Mystery of Edwin Drood* as there still is suspense as to the ending.

I do not wish to leave you with the impression that CSU is opposed to the assessment of teachers. Our plea is simply that universities be held responsible for assessing the competence of their students in the subjects they are to teach. We support a multifaceted assessment of teachers once they have enjoyed a solid residency and had the opportunity to develop their skills in teaching the subject at the grade-levels chosen—an approach consistent with proposals from the recent Carnegie Commission report.

I realize that our position flies in the face of the decision in many states based on the honest belief that a teacher test will create better teachers. My point is that a test *alone* cannot achieve this goal. In reform we need to be honest about all parts of the classroom-teacher milieu and work to strengthen each part. Class size, adequate nutrition for pupils, parental involvement, good texts and teaching materials are all a part of the picture. The teacher-in-preparation needs to develop interpersonal and presentational skills; the teacher must love working with young people, and needs training in psychology, behavior, that aforementioned course in tests and measurement, as well as good mastery of the subject to be taught. I am afraid we are now concentrating on testing only subject mastery as a panacea for all that concerns us in the classroom.

I share my worries on the teacher testing issue with you because I believe that you can and will provide good future solutions in this and other arenas where testing is looked to as a salvation. Let me broaden the glow from the crystal ball I see in three major areas of impact for education in the near future.

1. In my humble view, the issue of using testing actually to assist learning is the "hottest" and most important concept out there in higher education today—the buzz words are "value-added;" "assessing outcomes." You in this audience now have the capability to devise and to assist faculty in developing the appropriate instruments for this endeavor. Faculty, students, administrators are coming enthusiastically together in the pursuit of improving both teaching and learning. We're a little abashed in California for being slow to follow the prophet-in-our-midst, Sandy Astin, and get going. It's embarrassing but good for California to be out-trended by Tennessee, Missouri, New Jersey, Florida, and more. On October 15, we held a systemwide conference to kick-off our planning for systemwide efforts in student outcomes assessment.

We have, however, had for many years a large assessment-based program that has been successful. Students pursuing a degree in the Nursing Program of the Consortium of the California State University have the option of earning academic credit through non-traditional course work, through assessment or a combination of both. Students selecting the assessment option may utilize a variety of standardized examinations, including CLEP, DANTES, and ACT's PEP examinations. They may also utilize standardized exams developed by the University of the State of New York Regents External Degree in Nursing. In addition to taking written tests, nursing students may also take clinical performance tests involving either a role-playing client or a video format. While it is possible to complete the entire nursing curriculum by assessment, most students combine the instruction and assessment options to complete their degree requirements.

2. My colleagues and I look to you imploringly for help on the issue of cultural bias in testing. While I commend those of you who have and are working to eliminate the more obvious instances of cultural bias in standardized tests, I have the feeling that we have just scratched the surface in understanding the broader implications of cultural differences. It is incumbent on all of us engaged in disciplines that may lend further insights into the ramifications of cultural differences to assign high priority to developing a better grasp of these cultural dimensions. Although we have yet to discover the reasons, I am deeply concerned that the high failure rates of Hispanics and Blacks on tests in those 38 states that now test teachers threaten to diminish minority representation in the teacher force. Actually, we desperately need the opposite trend to provide good role models for our minority young people.

It is an interesting and challenging problem, and only you can solve it. I was fascinated to read recently in *Science* that the Minnesota Multiphasic Personality Inventory (MMPI)—it and I are about the same age—is now undergoing restandardization. "Normal" in that test was derived from a small sample of White, rural depression-era Minnesotans. (Does that make anyone abnormal who's not a fan of *Prairie Home Companion* and Garrison Keillor?)

It is a tribute, however, to the skill of McKinley and Hathaway, the fathers of the test, that it has been so enduring and so useful on a worldwide basis. The MMPI attests to the constancy of human personality and similar kinds of psychopathology, cutting across cultural lines.

Various indices such as the National Assessment of Educational Pro-

gress have provided generally accurate portrayals of levels of achievement in our public schools. One aspect of such test results disturbs me deeply and that is the consistently poorer performance of females in mathematics as compared to that of their male counterparts. The difference is far too great and has severe negative consequences for women, who now make up over one-half of students in higher education. Until this situation is remedied, women will not be represented as they should in the sciences, medicine, engineering, and even business. You have brought the problem to our attention; now we must collectively address it.

3. In my years as a bench scientist, the part I enjoyed most was "coming to grips with the data." All of you in this room have shared this experience. You sit with printouts and data books and reference charts and reprints, and if you're lucky, figure out what the results really mean. For me, that used to mean calculating how much calcitonin a fetal monkey released in response to a calcium stimulus. Now it means studying college-going rates and predicting baccalaureate productivity.

As a recent example, in the past we have always listed ethnic composition of our student body on a percentage basis. On that basis, Blacks dropped from 7.4 percent of CSU freshmen in 1981 to 6.7 percent in 1985—a dismal showing. However, in reality, in that same period of time, if counted in absolute numbers, the Black high school age population decreased 5 percent and Black CSU freshmen increased by 1 percent—an actual increase if you will in Black participation rates with the CSU. In California, the cohort of 18-year-old Asian Americans and Hispanics is growing, while White and Black populations drop. Thus, percentages mean little by themselves unless compared to absolute numbers.

Minority Enrollment Trends: California 1981-85

	HS-Age Population	HS Graduates	CSU Freshmen
Asian Americans	up 19%	up 47%	up 83%
Blacks	down 5%	down 8%	up 1%
Hispanics	up 1%	up 8%	up 31%
Whites	down 13%	down 15%	up 5%
TOTAL	down 7%	down 7%	up 12%

I urge you to use all of the considerable creative and analytic abilities you possess to come to grips with your data. I am convinced you can provide fresh insights, valid projections, early trends and even warnings

that will benefit his/her education. This process is also well-described by a caption at the Yakima Nation Museum concerning the crane:

> "The Crane stands immobile—but when he strikes he always comes up with a fish in his bill. The lesson is — patience should be followed by decisive action."

We need to study our data with patience and care but then act strongly on where they lead us.

> Father Guido Sarducci on "Saturday Night Live" recently described the "five-minute university." He points out that a few years after college, the graduate can tell all he or she remembers of classroom learning in five minutes. There is the five-second Economics 101 which is "supply and demand." Or how about the one-minute law degree based on "possession is nine-tenths of the law."

Please don't let us linger in the public's mind as the "five-minute university." In reality, we're depending on you to help us validate all that we're about, which is that much learning truly has occurred on our campuses.

The Why, What, and Who of Assessment: The College Perspective

JOHN W. CHANDLER
President, Association of American Colleges

The subject of assessment came to the fore in American higher education just two years ago when three major national reports on the quality of undergraduate education, appearing in quick succession, made it a major theme. The initial reaction of many presidents, deans, and professors was to dismiss assessment as the latest educational buzz word, with the expectation that it would go away quickly. When it became clear that assessment was not a passing fad, many persons in higher education tried to come to terms with it by claiming, perhaps somewhat defensively, that assessment was already a major part of what they did, and they wondered aloud what the fuss was all about. They pointed to their evaluation of student performance through tests, examinations, laboratory exercises, classroom discussions, and paper assignments. Faculty members cited the evaluations of their own performance by their departments, their students, and by faculty committees on appointments and promotions. This defensive reaction is still commonplace, but it is giving way gradually to a fuller understanding of assessment as it is practiced in industrial and military settings and in a few educational institutions where it receives special emphasis. This growing understanding of assessment is stimulating widespread interest in its potential for producing better students, better teachers, better courses, and better programs. The question is no longer *whether*, but *how*.

Much of the initial negative reaction to assessment was caused by the hoopla that surrounded its emergence as a major item on the agenda for the reform of undergraduate education. Many faculty members initially saw assessment as a threat because they were made to feel that it contained sacramental mysteries to which they were not privy and also because they saw assessment as a new set of demands being imposed upon them from without—from governors, legislatures, accrediting

associations, and from the administrations at their own institutions. They saw it as a threat to their freedom to run their courses as they saw fit and with no outside interference.

American higher education is distinctive and even unique in the degree of autonomy enjoyed by individual institutions. They make their own decisions about who is admitted, who teaches, how much faculty members are paid, and what students must do to graduate. The faculty, as the core of the university, cherishes its own special freedom, and the typical individual faculty member is vigilant against interference from outside the university as well as from non-academic components within the university.

The benign anarchy of the modern American university also has a student component. Some twenty years ago the curricular pendulum moved rapidly away from required courses towards greater elective freedom, thus giving students a larger measure of the kind of self-determination that their teachers enjoyed.

The major national reports on the improvement of undergraduate education, and especially the report of the Association of American Colleges entitled *Integrity in the College Curriculum*, charge that curricular incoherence is the result largely of the radical freedom of faculty members to teach what they like with little reference to the needs of students. And it is that alleged incoherence that the national reports are addressing. In the case of William J. Bennett's report, *To Reclaim a Legacy*, the prescription for restoring coherence, at least in the area of the humanities, is a core curriculum based upon the intellectual tradition of the Western world. The report by the Association of American Colleges, by contrast, proposes a scheme of order and purpose that is informed by nine intellectual, aesthetic, and philosophic experiences that nurture the characteristics of the liberally educated person. The AAC list of essential educational experiences and skills includes such items as inquiry and critical analysis; literacy, which embraces writing, reading, speaking, and listening; understanding numerical data; historical consciousness; scientific understanding; sensitivity to values; and aesthetic awareness.

Whether one uses the curricular model of Secretary Bennett, the curricular goals of the AAC report, or some other curricular rationale and design, assessment refers to the various procedures that are used to determine the extent to which individual students have met the curricular goals, mastered the prescribed subject matter, and acquired the skills and characteristics that certify them as having the essential marks of an educated person.

Assessment includes more than measuring the student's performance in reproducing or recognizing discrete items of information. Assessment focuses above all on exercises in which a student makes use of knowledge in new and imaginative ways to solve problems, raise fresh questions, and provide new insight. The highest educational achievement is literacy in the broad and deep sense in which the *Integrity* report defines it. No assessment scheme that relies excessively upon multiple-choice and short-answer tests is likely to lead to growth in students' literacy. A well-designed assessment program charts a student's growth in the grasp of the information and concepts in particular courses and programs. A successful assessment program provides feedback to both students and faculty so that both parties can judge their performance and make the necessary adjustments in their strategies and tactics as teachers and learners.

Alverno College is frequently cited as an institution—and perhaps it is the only institution—with a comprehensive assessment program that embraces all departments and programs and that stretches over the student's four years. Alverno students must demonstrate eight critical abilities, as they progress through six levels of performance in each of the eight abilities. In this educational slalom race, the student must make her way through more than 100 assessment gates.

Alverno's assessment program contains great merit, and dozens of institutions have taken valuable lessons from it. Alverno is remarkable in the degree of faculty collegiality and commitment that lie behind its assessment program. Still, very few colleges and universities are ready for a comprehensive program with a uniform methodology that stretches across all disciplines and programs. What they are much more ready for is a program that centers in individual academic departments and that focuses on the learning of students in the major field. Hence, most of the comments that follow refer to departmental faculties and programs.

The assessment movement holds considerable promise for encouraging faculties to exercise collective responsibility and to approach their educational tasks with a collegial mind-set. The department can be an especially effective unit for collective and collegial efforts to improve the quality of learning. The AAC *Integrity* report, in one of its strongest indictments of undergraduate education, charges that departmental major programs characteristically emphasize the number of courses required for a major but usually provide little or no rationale for the major and no compelling statement of the goals of the major. Consequently, the student is left to choose a specified number of courses from a large list but

is provided with little or no sense of the goals of the major and has little awareness of the particular knowledge and skills that one who majors in the field will possess. It is the *collective* responsibility of faculty members in specific departments to determine the goals of the major, design the courses and course sequences that will meet those goals, and devise exercises that will determine the extent to which major students are meeting the goals.

My sense is that in the past 25 years there has been a substantial decline in the influence and efficacy of the department as the locus of collaborative faculty work in the design and assessment of courses and programs. This decline is partly the result of the growing importance, in many institutions, of centers, institutes, and interdisciplinary programs. But the more fundamental change has been in the direction of greater individual faculty responsibility for designing courses and syllabi, and making and grading tests and examinations. Again, I have no statistical evidence, but my impression is that there are fewer team-taught courses now than formerly. In the natural sciences it is not uncommon to find a pyramidal structuring of the curriculum, with a logical progression from simple to more complex subject matter. But in the social sciences and humanities there has been a demise of introductory courses that serve as the prerequisites to advanced courses and which involve the collaboration of teams of faculty. Perhaps a single introductory course does not make sense in some disciplines and in some institutions. But I believe that departmental faculties should carefully discuss and examine that issue, along with the larger issue of the overall rationale for the list of courses offered and the relationships among those courses.

It is important to emphasize that assessment is not to be equated with testing and examinations. Tests and examinations are, however, important components of an assessment system. In the past quarter century there has been a substantial decline in the number of final examinations in courses, and comprehensive examinations are about as rare as the California condor. A well-constructed mid-term examination or final examination can be a valuable educational tool. It can provide essential feedback to both instructors and students. The review for such examinations can enable students to gain a comprehensive understanding of a course or a subject field and help them see relationships among the various elements of courses. As I recall my own teaching career, I believe that some of the most valuable investments of my time came when my colleagues and I spent long hours designing questions and exercises for final examinations in the introductory course and for the comprehensive departmental

examination, back before that examination was abandoned. Working together on those examinations compelled us to review the purposes and goals of particular courses and to consider the rationale of the overall structure of the departmental curriculum. Furthermore, those conversations were extremely valuable for young members of the department who were still making the transition from graduate student to full-time teacher.

One of the criticisms of American higher education is that the responsibility for instruction and for assessing student performance is vested in the same person, an arrangement that is contrary to practices in almost every other country. This practice confers enormous freedom and authority upon the American college professor. The model that prevails almost everywhere else in the world certifies the student on the basis of performance on examinations that are designed and graded by agencies and individuals who have had nothing to do with instructing the students. In recent years we have begun to see some modest movement in that direction in the American higher education scene. In a growing number of states, led by Florida and Tennessee, standardized tests are being used to determine not only which students are admitted to college but also which ones are admitted to advanced status and which ones are certified for graduation. The use of standardized tests holds great promise for elevating minimum standards of student performance. But if standardized tests assume too prominent a role in an institution, they can have a stultifying effect on teaching and learning. Such tests are not well suited for permitting a student to demonstrate his or her capacity for aesthetic judgment, critical thinking, moral sensibility, and other more subtle and elusive qualities of mind and character. Standardized tests do not permit students to demonstrate that they are literate in a meaningful sense. The best way to assess a student's learning and development, especially in the more advanced levels of the undergraduate experience, is through well-designed essay examinations, papers that are discussed before and after they are written, and well-run classroom discussions.

If assessment is to result in significant and lasting improvement in student achievement, assessment programs must remain under the control of faculties and not be imposed by legislatures and other external authorities. But to be credible and effective in the exercise of their responsibility for assessment, it is imperative that faculty members surrender some of their individual autonomy and work collaboratively. A modest step in that direction would be for faculty colleagues, both within and outside the same department, to work together in designing and grading one another's exams and in conducting oral examinations.

Another step could involve the use of examiners from outside the institution. Swarthmore College has used outside examiners in its honors programs since 1922, and the commitment there remains very strong. Wesleyan University uses outside examiners in its College of Letters and College of Social Sciences programs. The use of outside examiners is a long and strong tradition among the British universities.

In an attempt to promote more extensive experimentation with outside examiners, the Association of American Colleges is in the beginning stages of a program that is funded by a FIPSE grant. The AAC program will involve 18 colleges and universities grouped into six clusters of three each. Each cluster will contain three institutions of comparable size and character located in the same geographical region. Those three institutions will swap faculty examiners who will attempt to assess the level and quality of preparation of seniors in the major field. The experimental program will involve three disciplines in each institution. The senior majors will be examined by a team made up of faculty members from the other two institutions in the partnership. The form of the examinations, which will include both witten and oral components, will be determined by the examining teams in consultation with the departmental faculty members at the institutions in which the students are being examined. The use of the examination results will be determined by the home institution. The examination grade may appear as a separate item on the transcript. It may be used as a component in an honors grade or in a senior capstone course, if there is one. That question will be answered by each institution in its own way. What I wish to emphasize is that within the general structure of the program there is faculty control over the process and the substance of the exams. Our hope is that at the end of the three-year program, the participating institutions will have found enough value in the experiment to continue the program at their own expense, which should be relatively modest. We hope also, of course, that the success of the experiment will lead other institutions to adopt the same model. We believe that it holds great promise for improving the quality of learning, especially in the major.

While I have emphasized that departments should become more vital centers of collective responsibility for the quality of what is taught and learned, the collegial approach to these matters should also take forms that transcend departmental boundaries. I am encouraged by the increasing number of faculty workshops on specific educational questions. The improvement of student writing has become a goal for many institutions, and faculty workshops on writing have enlarged the capacity of faculty

members in all disciplinary fields to teach students to write better.

Faculty workshops and conferences that deal with teaching have also become more commonplace in recent years, and with good results. Few institutions have the resources to establish anything comparable to Harvard's Danforth Center, which provides an array of aids and services to faculty members who wish to become better teachers. But every college and university can call upon the experience and insight of its best teachers, almost all of whom are eager to tutor their younger and less skilled colleagues in designing and running courses and in developing methods for letting students use and demonstrate what they have learned. Presidents and deans demonstrate their commitment to teaching workshops by providing the resources and recognition that will encourage widespread participation.

What is encouraging about the assessment movement is that it is leading to promising developments in many different kinds of institutions. Alverno College, as I have previously noted, provides an exemplary model for assessing the growth of students in eight critical abilities over a period of time. Northeast Missouri State University, has, for the past dozen years, provided an instructive model of the "value added" approach to assessing higher education with the use of pre- and post-tests. The COFHE group of institutions, consisting of some of the nation's most prestigious research universities and liberal arts colleges, are giving careful and collective consideration to assessment. President Bok of Harvard has provided leadership among the COFHE group, and in his most recent annual report focuses in an exceptionally thoughtful way on the subject of assessment. He also addresses the assessment issue in his new book, *The Higher Learning*. Professor Richard Light of the Harvard faculty is conducting a seminar for members of the Harvard faculty and faculty members from other universities that is focusing on assessment in the context of relevant educational research findings. This is a badly needed emphasis, which points to the fact that very few college and university faculty members consult the relevant findings of learning theory and pedagogical research in designing and teaching their courses.

In summary, I wish to emphasize these points:

1. Assessment is not an end in itself but is, rather, a means to improving programs and achievement in learning.

2. A well-designed assessment program focuses upon the soundness and success of programs and courses fully as much as it emphasizes the progress of individual students. A successful assessment program

provides a variety of kinds of feedback to students, faculty members, and academic officers.

3. Assessment programs should be localized. That is, they should be geared to the goals of the particular institution, program, department, or course.

4. To be successful, an assessment program must be designed and run by faculty members, working collaboratively, with the use of outside resources (national standardized tests, outside examiners) that are relevant and useful.

5. While assessment usually involves testing, it is not to be equated with testing. Assessment of an individual student's learning goes beyond passive recall or recognition of particular items of information; it enables the student to use and apply knowledge in a range of exercises designed to demonstrate intellectual growth and versatility.

6. While minimum-competency testing may help to elevate threshold standards, the widespread use of standardized tests, especially when mandated by external authority, is a risky and unpromising means of improving the quality of programs and the overall educational achievement of students. Externally imposed assessment implicitly denigrates the role and authority of teachers; it can encourage them to design programs and courses to meet the norms of general tests rather than the higher learning they believe is most important. Over time, those trends would lead to institutional sameness and blandness.

The greatness of American higher education is its diversity, ingeniously developed by enterprising and imaginative professors, presidents, and deans to educate an enormous range of types of students and meet many different public needs. Constraining the freedom of faculty members and academic administrators to design programs that best meet the needs of their particular students would be a backward step for American higher education.

We are in the early stages of the assessment movement. It has some pitfalls, but the promise far outweighs the perils.

The Why, What, and Who of Assessment: The State Perspective

ELEANOR M. MCMAHON
Rhode Island Commissioner of Higher Education

As Gregory Anrig noted in his invitation to this conference, "The assessment of learning in college has emerged as an institutional, state, and national issue. It has been given particular urgency by legislatures in several states requiring evidence of student progress and of the standards being met on individual campuses." There thus seems to be abroad in the land significant support for Chaucer's maxim, "For that he naught assaieth, naught achieveth." I will seek, therefore, to address three questions: What has been the traditional role of the state in assessment in higher education? What appears to be the emerging role of the state? And what ought that role to be? Or, in Chaucer's terms and suggestive of the conclusion I will reach, "How ought the state to assaieth what it hopes its institutions of higher education will achieveth."

The Traditional Role of the State. As pointed out in the working document prepared for the ECS Commission on Effective State Action to Improve Undergraduate Education Agenda and Working Outline, 1986, of which I was a member, while the condition of undergraduate education is the subject of national concern, this concern is not without precedent. Undergraduate education was first subjected to national reexamination before the Civil War. The spotlight turned again toward higher education after World War II with the reports of the Truman Commission, the Carnegie Commission, and the ACE National Commission on Higher Education, all of which called for reforms in undergraduate education. However, it is important to note that these earlier reports, mindful of the strong American tradition of autonomy, were addressed primarily to institutions of higher education. They were not for the most part addressed to the states, to legislatures, or to state-level governing or coordinating boards—of which, indeed, there were few in what many would now undoubtedly view as "the good old days."

The Emerging Role of the State. Following an avalanche in the early '80s of national reports on elementary-secondary education, higher education moved into the limelight with a series of reports dealing, again and in particular, with undergraduate education. Central among these have been the NIE report "Involvement in Learning: Realizing the Potential of American Higher Education," the AAC report "Integrity in the College Curriculum: A Report to the Academic Community," the NEH report "To Reclaim a Legacy," and the ECS report "Transforming the State Role in Undergraduate Education." An unchanging part of the national report landscape is the fact that there was not before the Civil War and is not now a consensus on what improvements should be made. However, there is consensus on the importance of educational excellence at all levels for the cultural and social well being of all citizens. A latter day addition is the relationship seen between economic development and educational excellence. In fact, economic development is to education in the '80s what Sputnik was to education in the '50s.

While most of the reports that are stimulating current activity are national in their source (e.g., the National Institute of Education, the Association of American Colleges, the National Endowment for the Humanities), it is clear that at least one of the leading actors has changed—the call to action is increasingly directed toward the states. It is argued forcefully that this direction is appropriate as policy leadership, at least in terms of funding, shifts from the federal government to the states in a number of areas including education.

Peter Ewell in his paper on "Levers for Change: The Role of State Government in Improving the Quality of Postsecondary Education" (1985), raises the question of why state government should get involved. His response is that calls for state involvement rest on two arguments: the considerable investment that most states make in their systems of higher education and the demonstrable connection between the effectiveness of such systems and the fulfillment of other state objectives such as economic development.

A 1985 College Board Study (Goerty and Johnson) showed that indeed the states are involved in what are perceived as qualitative initiatives: at that point in time 24 states had set minimum admission requirements for freshmen at all of their public institutions and a number of other states were in the process of considering such actions.

Next in the order of magnitude, clearly expanding, and closer to the quantitative mode traditionally associated with assessment, have been state-mandated achievement tests. By way of illustration, at the entry

level the New Jersey Board of Higher Education has developed a College Basic Skills Placement Test Program that covers reading, writing and mathematics and is administered to all students entering New Jersey public colleges and universities as well as 11 private institutions in the state. At the lower division matriculation level is Florida's College Level Academic Skills Test (CLAST), which requires that all students in public institutions attain a specified score as a condition for either receipt of an associate degree or enrollment in upper-division courses. The test is also required of any student in a private institution who receives state financial aid. In addition to the state-level minimum competency examination, Florida has established specific curricula standards that require all students to complete 12 semester hours of English, including 6000 words in writing in each course and six semester hours of mathematics. The state of Georgia also requires a state-wide "rising junior exam" and similar tests to determine readiness for upper division work are under consideration in a number of other states. Finally, some states have required examinations for particular fields of study, most notably teacher education. Mississippi, for example, has set a minimum score requirement on the ACT/COMP exam for those interested in entering teacher education, and a number of states have analogous standards for the SAT. As far as graduation tests are concerned, few states have taken action in this area, a notable exception being Georgia, which for some 13 years has required students to pass a Regents' Exam in order to receive a degree.

Other state-level initiatives are in process. The Maryland State Board of Education conducts a comprehensive student follow-up study, and a number of states, including Rhode Island, have mandated a strengthened program review process at the institutional level. Some states have also revised their funding formulas to reward colleges that demonstrate gains in student learning. Illustrative of this latter approach is Tennessee's Performance Funding Program. Finally, a number of states such as Virginia and Connecticut have established competitive grant programs aimed at improving instruction.

In a recent ECS survey of state initiatives designed to improve the quality of undergraduate education, Kozloff (1985) found that states had either taken or were contemplating action in 12 basic areas. The seven most prevalent categories of activity, each engaged in by at least 10 states, were improved articulation between the elementary-secondary and postsecondary levels, the establishment of policies to govern cyclical program reviews, incentive funding, revision of teacher education curricula, increased admission standards, assessment of student performance—

most typically in relationship to basic skills—and development and strengthening of the general education core.

A review of selected state plans makes it clear that while virtually every state is involved in some type of "quality improvement" activity, the nature of state involvement varies greatly, ranging from rather low-key decentralized programs that rely largely on incentives to stimulate institutional action to comprehensive and highly prescriptive regulatory programs, with all kinds of intermediate variations. This leads quite naturally to the question of the appropriate role for the state.

What Should the Role of the State Be? It is the position of Ewell in the paper cited above that, in pursuit of quality, the role of the state must be as strong a one as it has been in elementary-secondary education. But Ewell then goes on to make an important distinction. He notes that higher education is different both in terms of the problems it faces and its governance structure. Ewell sees no lack of basic talent in college classrooms, nor does he see colleges and universities dealing for the most part, despite some problems of remediation, with highly deficient student populations. While data from statewide basic skills tests would suggest that Ewell is overly complacent about the skills of college students, his second point is more persuasive. He argues that, historically, colleges and universities have been decentralized and largely self-governing enterprises founded on principles of individual faculty authority and academic freedom. Thus, he maintains, colleges and universities have considerable capacity for self-improvement and indeed that if such an improvement is to last, it must come from within. This he sees as presenting a paradox to external authority and, in particular, to the question of the state's role in assessment. More specifically he says:

> On the one hand, if higher education is left entirely to itself, it is likely to neglect socially important tasks. On the other hand, if state regulatory authority is applied directly, the very mechanism for effectively achieving these socially important tasks may be threatened (1985).

Ewell concludes, and I would agree, that unlike the situation in elementary and secondary education, the state role in improving the quality of higher education, while significant and essential, must at the same time and for the most part be relatively indirect and circumscribed. The key is "to develop policy mechanisms which trigger institution-level efforts towards self-improvement."

One notable exception to the indirect approach recommended by Ewell is at the system level, where the objectives for higher education must be broader than the sum total of individual institutional objectives. The system must be concerned, for example, with questions of access at all levels, with the availability of different types of institutions, with the efficiency of programmatic distribution, with transferability, and with applied research, which serves statewide economic development needs and which is likely to go beyond the capacity of any single institution. Thus, Ewell would see the state as having the primary responsibility for defining the specific terms of systemic effectiveness. For Ewell such a perspective is important to ensure that the drive to quality does not induce institutions automatically to become more like one another and thereby fail to serve the differentiated objectives that must exist at the state level.

Ewell's approach is reinforced and extended by Harvard's president, Derek Bok. In an address given at the centennial celebration of the New England Regional Accreditation Association, Bok (1985) pointed out that one of the distinguishing features of American higher education is its remarkable freedom from government control. Typical characteristics, which, as Bok notes, are familiar to the point of seeming too obvious to mention, are the appointment of new professors without government review—even in public institutions, selection of students, determination of curricula, and considerable discretion if not complete freedom in the allocation of funds. This tradition of freedom and autonomy is very different from the pattern of continental Western Europe where, for example, admission is usually guaranteed on passing of a state, as distinguished from an institutional, examination, where, while institutions recommend faculty appointments, the state ultimately determines who will fill academic posts, where there is little discretion in the allocation of funds, and where, in general, only in matters of curriculum and course content do institutions have freedom analogous to that of institutions in the United States. It is Bok's conclusion that our decentralized and largely competitive system makes our universities more venturesome, more variegated, and more adaptable to changing needs, all characteristics extremely important in a period in which the United States is concerned about its international competitive position.

While Bok concludes that competition has served well the qualitative interests of society in higher education, he is not complacent. He sees competition resulting in quality in a number of areas, such as sports, where effectiveness can be measured, but he does not see the same process at

work in determining the quality of academic programs. Neither does he see outside pressure as consistently effective. One of the most interesting and valuable insights in the analysis to which these observations lead him comes from his examination of the impact of the Flexner report. That report issued in 1910 resulted in a major reform in medical education. However, there were significant portions of the report that had virtually no influence, sections concerned with studying the psychological and social dimensions of illness, for example. The factor which in Bok's view distinguished the influential portion of the Flexner report from those portions which were without influence was that the former conformed to values widely held within the faculties of the better schools while the latter did not. Bok thus provides historical evidence in support of Ewell's point on faculty involvement and consensus as a necessary condition of reform.

As far as the effectiveness of external funding in pursuit of quality is concerned, Bok notes that such funding has served as a catalyst in higher education, but generally in the area of research. When it moves into the area of quality of instruction, he is less optimistic about its impact. This funding can serve as a stimulus, but once again it is the conclusion of the twenty-fifth President of Harvard that serious reform will only emerge "through the combined efforts of a determined administration and a willing faculty."

These analyses of Ewell and Bok suggest a number of principles that should guide state action. First of all, because of its investment, because of the existence of state objectives beyond institutional ones, and because of the responsibility which the state has to be accountable to the public, there is an important role for the state in quality improvement and in assessment. However, it is not a monolithic role. There is as much variation in higher education between states as there is within states, so that the notion of a single model of state involvement is simplistic. While there are broad commonalities, the number and range of institutions, the pattern of state governance and coordination, the resource investment, the mechanisms of assessment already in place at the institutional and state levels, and the desired outcomes vary from state to state. The nature of state involvement should be driven by these widely variant factors and therefore there are and must continue to be a variety of assessment models. These models must preserve the variety and richness that characterizes different systems (as well as institutions) of higher education, and they will do so only to the extent that they are custom-tailored to the needs of a particular state. The New Jersey assessment model is, for

example, quite different from the Virginia model, and the rationale for each is persuasive. If the price of quality assessment is homogenization at either the system or institutional level, then it is likely to diminish rather than enhance quality.

Secondly, while the state's involvement in system-level assessment should be direct and prescriptive, the state's involvement in institutional assessment should be relatively indirect: the state should define the general categories and, in some instances, the parameters of assessment and should ensure through an appropriate monitoring process that assessment takes place, but it should leave most, if not all, of the particulars of assessment to individual institutions. I say "most, if not all," because in some instances, particularly at the entry level, particularly in relationship to basic skills, and providing there is allowance in the standards set for legitimate differences in institutional selectivity, development of state-level standardized tests may well make sense in terms of both efficiency and effectiveness. On the other side of the argument is the potential leveling influence of state-mandated tests, particularly if they go beyond basic skills, and the strong possibility that such tests will stifle what might be more diverse and more appropriate institutional assessments.

The principle of indirect state involvement suggests a number of corollaries. The state should use incentive funds as the primary mechanism for encouraging institutional administrators and faculties to develop effective assessment programs. In developing standards for the distribution of incentive funds, emphasis should be placed, first of all, on faculty involvement and secondly, in recognition of the diversity and complexity of the undergraduate experience and the concomitant necessity for complex assessment strategies, on the use of multiple methods of institutional assessment. One of the concerns frequently expressed about the current assessment movement is the tendency to look for a single measure of quality analogous to a simple-minded reliance on a single SAT score as a measure of the comparative quality of a state's elementary and secondary schools. A second concern is that assessment in higher education will be limited to basic skills. Undoubtedly these are easier to measure than are higher-order skills and knowledge. But to fail to measure the latter would be to ignore the central purposes of higher education. Valid and reliable assessment measures will be as diverse and as wide-ranging as are the programs and institutions under review. They will recognize that assessment is more than testing, that it must go beyond the measurement of minimum competencies, and that, while it will certainly rely on data, it will rely equally on informed professional judgements. Most

importantly, since effective assessment is not an end in itself but rather a means to reform, state incentive programs should encourage institutions to develop assessment programs that are largely formative and that therefore include mechanisms such as faculty development for translating assessment information into improved practices.

Finally, there is a point made by Bok that is central to the assessment drive if it is to result in quality improvement, and that is our limited technical capacity to recognize quality and to compare the effectiveness of alternative programs. Bok sees this as requiring a major research effort on the part of our universities directed toward developing better ways of measuring the impact of undergraduate education and assessing the various methods of instruction. In other words, the state of the art is relatively primitive, and resources from the state as well as the federal government and private foundations must be invested and the research capacity of our institutions engaged in developing our presently limited repertoire of effective methods of college level assessment.

I would add just one footnote to Bok's point and that is that, while the state of the art is relatively primitive, a lot more quantitative and qualitative data is available than is used effectively. Thus, while research is needed, existing data is extensive and to the extent it can be incorporated in a sound assessment package, it should be used. In many instances, what is needed is not so much the generation of new data as the extension and translation of existing data into useful information.

The Role of the State in Two National Reports. The thrust of the ECS report (1986) to which I have already referred is consistent with the Bok and Ewell positions and with the principles suggested: it explicitly recognizes the growing evidence of critical connections between the quality of higher education and regional economic development; it recognizes the importance of meeting the needs of an increasingly diverse student population; it emphasizes the assessment of both student and institutional performance; and most relevant to the topic before us, it sees the state as having major responsibility for meeting these challenges while simultaneously recognizing the primacy of the institutional component.

The ECS report defines eight challenges, which are then translated into 22 recommendations. The organization and statement of these recommendations clearly recognizes both the responsibility of the state to develop a comprehensive state strategy for educational excellence and the general principle of subsidiarity. It sees the role of the state as mover and shaker and the role of the institution as designer and doer. It thus

suggests that while most of the action to enhance quality must be conceptualized and carried out at the institutional level, over and above the institutional responsibility is the responsibility of the state to stimulate that action, to ensure that it takes place, and to add to it assessment directed toward broader state responsibilities across the entire system. Finally, the ECS report urges states and institutions in evaluating undergraduate education not to stop at assessment but rather to recognize that its purpose is to improve teaching and learning. The report concludes that the ultimate "test" is the extent to which the results of assessment are incorporated into an institution's strategy to improve teaching and learning and a state's strategy to improve its system of higher education.

In the recent report of the nation's governors entitled "Time for Results: Governors' 1991 Report on Education" (1986), significant state action is also called for, but again in terms consistent with the Ewell and Bok postures. Specifically, governors, state legislatures, state coordinating boards, and institutional governing bodies are asked to ensure the existence of a clear definition of the role and mission of each public higher education institution in their states. These same bodies are asked to give high priority to undergraduate instruction in all of their institutions and most particularly in universities that normally give high priority to research and graduate instruction. Further, each college and university is asked to implement systematic programs that use multiple measures to assess undergraduate learning, while states are urged to provide incentives for the development and institutionalization of such programs.

A Modest Beginning. In Rhode Island we have tried to adhere to the principles suggested. In the context of a "Master Plan for Quality," the Board of Governors has developed policy mechanisms that are designed to stimulate institution-level efforts toward self-improvement. The principle of subsidiarity has been applied.

More specifically, in consultation with the institutions of higher education, two general courses of action are being pursued: all programs are on a cyclical schedule of either institutional level review or national program accreditation, and assessments from these reviews will be shared annually with a subcommittee of the Board of Governors, and a set of "indicators of quality" has been defined in generic terms. These range from admissions standards and retention and completion rates to value-added measures and placement data. As a set, they include both qualitative and quantitative as well as input and output variables. Each institution must report to the Board every three years on the defined indicators of quality

as these indicators fit the particular institution. The report on trends in admission standards coming from Rhode Island's research university would thus be quite different in content and focus from an analogous report from its open access community college. The basic approach is to use existing data wherever possible but to reorganize that data so that it becomes a part of a comprehensive assessment package.

As a stimulus to further institutional action, the Rhode Island Board of Governors has proposed to the Governor and the Legislature, in the context of its FY '88 budget request, financial support for an "Incentive Fund for Excellence" which, if approved, would be used to finance programs designed to improve quality at the undergraduate level.

In Summary. The traditional role of the state in higher education assessment has been minimal. That role is clearly expanding as both the interest of the state and its investment in higher education expands. The current patterns of state involvement vary widely from indirect incentive programs to direct regulatory action. While the state has primary and direct responsibility for the assessment of the system of higher education, the task of institutional assessment and resultant reform must be required but also stimulated and rewarded by the state and designed and carried out by the faculty and administration of each institution. As a result, patterns of institutional assessment will vary, as do existing patterns of state assessment. However, in both instances there should be significant formative as well as summative elements. Finally, the state should also recognize that an essential resource in improving the quality of assessment is the research capacity of its institutions of higher education and it should therefore engage in direct funding of research on assessment.

The ultimate principle underlying the assessment drive should be that whatever the pattern of state assessment, it must preserve and enhance the richness and variety of American higher education. This suggests that if the hand of the state is to be iron, the glove must indeed be velvet or, returning to Chaucer, while the state that naught assaieth, naught achieveth, the state that assaieth alone or over-assaieth, under-achieveth.

References

AAC Report. "Integrity in the College Curriculum: A Report to the Academic Community." The findings and recommendations of the Project on Redefining the Meaning and Purpose of Baccalaureate Degrees. Washington, DC: Association of American Colleges, February 1985.

Bok, Derek C. "The Quality of Education in American Universities." *Education and the Welfare of the Republic.* Boston, MA: New England Association of School and Colleges, Inc., December 1985.

Education Commission of the States. "Agenda and Working Outline: Working Party on Effective State Action to Improve Undergraduate Education." Denver, CO: ECS Headquarters Office, April 30, 1986.

ECS Report. "Transforming the State Role in Undergraduate Education." The Report of the Working Party on Effective State Action to Improve Undergraduate Education. Denver, CO: Education Commission of the States, July 1986.

Ewell, Peter T. "Levers for Change: The Role of State Government in Improving the Quality of Post-secondary Education." Denver, CO: National Center for Higher Education Management Systems (NCHEMS), November 1985.

Goertz, Margaret E., and Linda M. Johnson. "State Policies for Admission to Higher Education." New York: College Entrance Examination Board, 1985.

Kozloff, Jessica. "Survey of Recent State Initiatives to Improve the Quality of Undergraduate Education." Denver, Colorado: Education Commission of the States, December, 1985.

NEH Report. "To Reclaim a Legacy." A report written by William J. Bennett, Chairman of the National Endowment for the Humanities and based on the finding of the Study Group on the State of Learning in the Humanities in Higher Education, November 1984.

NGA Report: "Time for Results: The Governors' 1991 Report on Education." Washington, DC: National Governors' Association, August 1986.

NIE Report: "Involvement in Learning: Realizing the Potential of American Higher Education." Washington, DC: National Institute of Education Study Group on the Conditions of Excellence in American Higher Education, National Institute of Education, 1984.

The Why, What and Who of Assessment: The Accrediting Association Perspective

THURSTON E. MANNING
Director, Commission on Institutions of Higher Education
North Central Association of Colleges and Schools

As a former physicist I have a strong longing to begin with a clear and confining definition of the term "assessment." But as a present pragmatist—as well as a linguistic empiricist—I cannot bear to constrain a word that has come to have so many meanings to so many people. The fact is that this poor word has been pressed into service as a prestigious substitute for everything from the plain vanilla of "testing," to the pistachio walnut rum raisin of "a multidimensional process of judging the individual in action."[1]

So let us agree that "assessment" will mean any activity, from the simplest to the most complicated, directed at reaching a judgement. That won't please the purists among us anymore than it pleases me; but it may not be all bad, as an anecdote may illustrate:

I cannot now remember where I saw it, but long ago I read a comment of Percy William Bridgman, the distinguished physicist. Bridgman, who as a promulgator of operationalism had quite a reputation as a philosopher of science, was once asked how he defined "the scientific method." Undoubtedly annoyed at the neat recipes popular in elementary textbooks, Bridgman replied, "The scientific method is doing your damndest." Since the corpus of Bridgman's work shows clearly that he paid full attention to the doctrines of reproducible experiments, theories based on valid data, and full public disclosure of mistakes—all those ingredients of the textbook definitions of "scientific method"—the remark may be as puzzling as it is memorable. But I think it illustrates a distinction helpful in discussing assessment: the distinction between doing and validating. A Bridgman doesn't have a set of rigid guidelines to tell him in advance how

to do an experiment; in the doing he "does his damndest"; afterward he is concerned whether what he has done meets necessary criteria. Bridgman was saying that "the scientific method"—despite its name—is not a method we use to do scientific work; it is a way of telling whether what we have done is scientific work. Not many textbooks make that distinction, and many an eager student, searching for rules to guide his or her work, has been misled.

So with assessment: the question burning in our colleges is "How to do it?" Our temptation is to provide a list of rules and an "assessment method." Alas, rules may be helpful in telling whether what has been done is assessment, but they are misleading and frustrating as marching orders for novice assessors. Assessment (or evaluation) theory is still embryonic, and the most we can do is to teach how to identify available instruments (much as the science teacher identifies instruments and techniques) and provide case studies (as the science teacher provides papers telling how someone else did something similar). To do assessment you "do your damndest": testing here, interviewing there, using results separately, or mixing everything together. We have some ways of telling whether the result is right or not, but so far no one has a way of telling what to do. With that observation, unhelpful as it may be, I proceed to Why, What and Who.

Why

Why assess the outcomes of higher education? The classic answer says we do it for two reasons: to find out what has been accomplished, and to find out how we might accomplish it better. Accrediting associations are interested in both reasons.

Accrediting associations assess institutions and programs in part to find out what they have accomplished. If the result is that they have accomplished at least an acceptable level of excellence, the institution or program passes one—but only one—criterion for accreditation. To say this in another way: accreditation of an institution or program rests in part on the assessment of institutional or program outcomes. That fact surprises some people, who have a vision of accreditation as concerned exclusively with narrow requirements of resources and processes. That accreditors make judgements based *exclusively* on the number of books in the library, the number of full-time faculty members, and the distribution of hours between class and laboratory is an idea that dies hard.

But die it did: formally about 50 years ago; actually perhaps 10 years ago; virtually ever since the beginning of postsecondary accreditation some 70 years ago. One reason is that requirements of resources and processes at best provide necessary conditions for education, but not sufficient ones: books in a library may be necessary if a student is to study; but books in a library do not ensure that a student does study.

Given the identification of accreditation with the requirements of resources and processes, it may be startling to observe that such requirements may not be even necessary. Take as an example requiring full-time faculty members: why do you need full-time faculty members to have acceptable educational quality? Once you say why they're needed, consider whether there are alternative ways to meet those needs. There are such alternatives, and some institutions have used them. If an institution can show that its graduates, taught by part-time instructors, perform as well as graduates of an institution using full-time instructors, the institution is showing that an accrediting requirement of full-time faculty is not necessary for educational quality. The test needs to be not just what the institution has, but also what it does. That means looking at outcomes. Here is another reason for assessing outcomes: it can be a way to demonstrate the success of alternative means. Outcomes assessment can open the door to fruitful alternatives in educational procedures.

Outcomes assessment can also help identify where things might be done better, and accrediting associations have from the beginning been concerned with institutional and program improvement. At the most elementary level, disappointing outcomes raise the question, "Could we do better?" Sophisticated outcome assessments help to identify where we could do better. This is, for example, a primary use of outcomes measures at Northeast Missouri State University.[2]

Accrediting associations do not place exclusive reliance on outcomes. An obvious reason is the undeveloped state of outcome measures; even institutions that have devoted much effort to assessing outcomes appear to improvise much of the time, and the most ardent advocates of outcome assessment acknowledge that much work needs to be done.

A more fundamental reason, which makes clear that assessing outcomes cannot be the exclusive means of judging institutions or programs, is that outcome measures are by their very nature retrospective. That is, they can at most tell us what has happened in the past; they cannot assess current conditions or give an estimate of future success or failure. That characteristic of outcome measures seems to me obvious, but I don't find it emphasized in the literature. Yet obvious it is: the successful graduates

of the Yale Law School reflect what the School was, not what it is; after all, the graduates are not there now. Only by combining information about outcomes with knowledge of how they were achieved and of how present resources and processes compare with those of the past can we reach valid conclusions about the present state of an institution. You will note that in making this criticism of outcomes assessment I am not discarding it; I am merely saying that the questions we wish to answer in accreditation need more information than outcomes alone can provide.

What

What should be assessed? We need to be clear about two things: the object of assessment (students, institution, program) and the characteristics to be assessed. Accrediting associations assess institutions and programs, but they are not indifferent to assessment of students. The reason is that the doctrine of accreditation says that institutions and programs are to be assessed against their stated (and acceptable) purposes. Among those purposes for educational institutions must be goals for the educational achievements of their students. Thus assessing whether an institution or program is achieving its purposes includes whether its students are achieving satisfactory educational goals.

Educational outcomes are, of course, only part of the outcomes of higher education, and our assessment cannot be restricted to them. Accrediting associations ask institutions to provide outcome assessment also for their research activities and their work that falls under the rubric of "public service." Here assessment mechanisms are very rudimentary, partly because these activities loom large in only a few institutions, but mostly, I suspect, because no one has put much thought into how to do it.

Underlying the assessment of outcomes are the purposes an institution aims to achieve, and one of the greatest problems institutions have is stating clearly what they wish to accomplish. Without clarity of purpose it is impossible to judge whether purposes have been achieved; if you don't know what outcomes you want, you can't decide what outcomes to assess. The current emphasis on outcome assessment has helped the accrediting associations in their efforts to have institutions make clear what their purposes are.

Assessing the outcomes of higher education thus brings a focus on the goals of institutions of higher education: Are they worthy? Are they

appropriate? Are they acceptable? That suggests that the goals themselves need to be assessed—assessment upon assessment. Accrediting associations also do this, albeit more informally, using general ideas of social acceptability as the criteria for judging institutional and program goals. But in any accrediting assessment of goals lies an underlying concern: that the educational outcomes of an institution be appropriate to the credential it confers. That implies an understanding of the meaning of a degree, and in the heterogeneity of American higher education that understanding is unexpressed, being close to the idea that "I can't tell you what it is, but I know it when I see it." Perhaps the assessment movement will help define what a bachelor's degree is (to say nothing of a doctor's degree). I doubt that it will, and that raises an interesting point: if we can't say what achievements identify the holder of a degree, how can we assess whether the holder has accomplished those achievements? Assessing outcomes follows from knowing what outcomes are desired. At a fundamental point—defining the credentials we confer—we have no national agreement.

Who

And so we come to who should assess the outcomes of higher education. The accrediting associations have a clear party line on this one: both you and we should do it.

Who are you? In our view, you are any interested party, but especially the institution itself. The concern of the accrediting associations is not only with judging whether an institution or program is of acceptable quality, but—and increasingly—with institutional improvement. The emphasis on self-assessment has grown as the accrediting associations have learned that self-assessment is an effective means not only to identify areas needing improvement but also to convince those who must make the improvements that they should make them. Higher education may be populated with grown up men and women, but they still have the peculiarities of human beings. The private assessment of institutional weaknesses is more effective than the exhortation of the accrediting association in effecting improvements—just as the private assessment of our corpulent self in the bedroom mirror is more effective than the directive of our physician in inducing weight loss.

There are other players in the game—state agencies, for example—and they too should join in assessing the outcomes of higher education.

But note that the different players may have somewhat different ideas about what assessment results are needed. Accrediting associations ask about merit: whether an institution or program is operating at an acceptable level of quality. An institution may ask also about efficiency: whether quality can be maintained with fewer resources. A state agency may ask about worth: whether the state needs two accredited schools, or whether one could be closed without significant social loss. The who of assessment can affect the what of assessment.

There is a fundamental concern of all parties for the quality of the educational enterprise, suggesting that a sharing of assessments would be advantageous. Yet surprisingly such sharing is not common. Accrediting associations find that institutions and programs like to start out fresh to conduct the self-assessment and gather the data the accreditors ask for. Sometimes there seems to be the conviction that we want it that way, that work already done just can't be pertinent. That isn't true—good, continuing institutional assessment is just what accreditation needs, and I can say from personal experience that the continuing assessment programs of such institutions as Northeast Missouri State and Alverno College mesh well with accrediting needs.

Why do we tend to keep assessments in separate compartments? The problem seems to me to be related to our emotions about assessment. That portly fellow in the mirror doesn't want to wear a sign that says, "Fat"; he's willing to drop a pound or two, but it has to be done quietly. Many an institution is fearful that public knowledge of the "concerns" of its accrediting association will be translated into enrollment losses if not worse; it wants to improve, but wants to do it quietly. Assessment of outcomes carries with it the possibility that the outcomes will not be what we would like; who does the assessment can strongly affect the public disclosure of distasteful results. That, in turn, can lead to attempts to control disclosure if not the assessment process itself. The validity of assessment results often depends on the cooperation of those who are assessed; even doing one's damndest won't work if the object of assessment is too uncooperative. Valid assessment of outcomes requires sensitivity to the human and social context; it is not just routine application of instruments or techniques.

Cautions and Improvements

From the perspective of accrediting associations we assess outcomes of

higher education to find out how well goals have been achieved, and, by combining with other information, to suggest how achievement can be improved. We assess institutions and programs against the achievement of their worth goals; and we assess those goals against social acceptability. We urge that institutions and programs themselves do such assessment along with accrediting agencies and other parties at interest. That sounds sensible, and I would be surprised if anyone objected to it.

Another piece of accreditation policy is that whatever is done could be done better. So I can't say that the assessment of the outcomes of higher education needs no improvement: that would be contrary to the dogma promulgated by the organization that puts the food on my table. Besides, I happen to think that we really could do better. Why don't we?

Fundamentally the reason is that in higher education we deal with complex human students immersed in complex social organizations. That results in complex goals and correspondingly complex outcomes. Even getting agreement on goals can be difficult. To illustrate, consider this goal: every college graduate should be prepared to move smoothly into immediate employment. Sounds good, but we don't agree on that goal even for some apparently obvious cases; if we did, we would withdraw accreditation from every law school that had a graduate fail the bar examination, since you can't practice until you pass—but that would mean that every law school would lose accreditation, and probably Harvard would object. Complex goals and complex outcomes make assessment difficult.

Even more difficult is making assessment of outcomes serve institutional improvement. What we know—and our friends at Northeast Missouri and Alverno certainly agree—is that such use of outcomes assessment is hideously difficult and probably so strongly dependent on the institution that case studies cannot serve as models. They are examples of what can be done by doing one's damndest and illustrate well that doing assessment is different from knowing that an assessment is valid.

When tasks are difficult we sometimes become impatient and take shortcuts. The current attention to assessment of outcomes seems to me to be subject to abuse as some, often with the best will in the world, undertake assessment because it is fashionable, pressing into service whatever assessment instruments are available, whether or not they are appropriate to the purpose to be served. Because it is easier to use an assessment instrument than it is to agree on desirable goals and devise appropriate assessment of outcomes toward those goals, we may well see misuse of existing tests and other measures. Indeed, we have already seen

the SAT taken in some quarters as a definitive measure of scholastic achievement. A particular danger here, it seems to me, is misguided legislative zeal in adopting statewide testing programs without differentiation among institutions within the state.

To help with the difficult task of outcomes assessment we need more and better tools and procedures; these are the analogues of the scientist's instruments and techniques. Here organizations like ETS and ACT can serve as instrument makers, devising ways by which various outcomes might be assessed. We also need examples: these are the analogues of the scientist's literature describing the work of others. Here we need better ways to share case studies of institutions that are doing their damndest to assess outcomes; most of that information is now buried in ephemeral newsletters and internal documents. Without better tools and procedures, without more examples to suggest ideas and provide the caution of failures, we will continue to fumble. I understand present fumbling — we have never learned how to do something new without making mistakes — but we must lower the incidence of fumbling. That is why we must be concerned not only with the assessment of outcomes but also with the improvement of the assessment of outcomes.

Accrediting associations have been promoting outcomes assessment for a long time; we believe that it is an essential part of certifying institutional quality and promoting institutional improvement. It is comforting to have others help push with us for outcomes assessment. I hope that all our pushing can move higher education toward more and better assessment of outcomes, both to demonstrate that higher education is doing what it should do and to help it do its work better.

Footnotes

1. Loacker, et al. "Assessment in Higher Education: To Serve the Learner," in Adelman (ed.), *Assessment in American Higher Education*. Washington: U.S. Department of Education, 1986

2. *In Pursuit of Degrees with Integrity: A Value Added Approach to Undergraduate Assessment*. Washington, American Association of State Colleges and Universities, 1984. A shorter description is to be found in Ewell (ed.) *Assessing Educational Outcomes*, San Francisco: Jossey-Bass, 1985.

Critical Validity Issues in the Methodology of Higher Education Assessment

Eva L. Baker
UCLA Center for Student Testing, Evaluation, and Standards

Validity is the grand old concept of assessment. It stands for a complex set of ideas involving the purposes of assessment, the match of information obtained to such purposes, and the process by which information is verified. Validity in testing, as in English, is about truth. This paper focuses on increasing the validity of student assessment in higher education.

Since validity is an apparent good, why do we have a problem with it in higher education—or anywhere? Our validity problems occur because we frequently are unclear about the purpose we are serving with our assessments, a situation that also clouds the inferences we should make from our findings.

Traditionally, at the postsecondary level, we have tested students for admission and placement. Admissions testing has drawn public attention because of its centrality in the allocation of equal educational opportunity and because the average admission test score has become a shorthand description for the educational standards of colleges and universities—the purported goodness of the education directly related to the difficulty of admission. More recently, average admissions test score has been applied, in a similar way, to evaluate the precollegiate educational effort. Although it has been common at private schools to judge educational quality in terms of the number of students admitted to the most elite postsecondary institutions, it was only relatively recently that such college admission test scores were used to compare state educational systems (U.S. Department of Education, 1985). Both uses of admissions tests raise obvious problems relating to the validity of inferences: are we talking about the quality of the educational institutions themselves, the quality of their clients, or some unknown combinations of the two? Furthermore, such quantitative shorthand whets some appetites for other

simplified measures of educational quality. So, of increasing interest to the postsecondary community and those compelled to comment about its effectiveness, is the utility of student achievement measures for assessing postsecondary educational quality. Driving these interests in student assessment are legitimate public concerns about higher education costs and benefits. The spate of attention to this issue by the Federal establishment was perfectly predictable: as precollegiate educational programs were shifted to States for management, the majority of the remaining federal educational investment was directed to postsecondary students. Accountability went to college.

Present Methods

From all reports, each of the existing systematic assessments of student academic performance in colleges and universities has developed through top-down mandate. How high up that top is varied, with the present ceiling at the statehouse. The intended purposes served by such mandated student assessment include accountability (reporting to legislatures), certification (verifying performance for existing teachers), or institutional self-study (McClain and Krueger, 1985). Although assessment systems may begin with one ostensible purpose (who goes to what segment of higher education), a mutation such as outcome assessment is not hard to imagine. A major fact about testing is that whatever its original purpose, the findings from assessment are always used for something else.

From all appearances, many existing assessments of postsecondary students share the methodology and flavor of precollegiate, large-scale testing activities. The measures are standardized. They are formulated for and administered to the group. They often focus on minimums. They have great symbolic value, and their functional value is unknown. To the extent that student assessment measures become widespread, I will predict that their original purposes will be transformed and that they will also drive out other indicators used to evaluate comprehensively the quality of higher education institutions. Simply look at precollegiate education as relevant history. Mandated, large-scale testing occurred because the precollegiate system had no convincing information about its quality. No information was available to refute claims that kids couldn't read and write, let alone do fractions and analyze Shakespeare.

Assessment as an Instrument of Educational Reform

At the heart of this discussion is the use of assessment as a bureaucratic tool. Bureaucracies seem to see a formal assessment program serving at least two purposes: first as an indicator of system quality; second, and increasingly more importantly, as an intervention. In precollegiate education, for instance, imposition of mandated testing is seen in itself as a major educational reform rather than as a way to measure the effects of changes in educational services. Testing is a classic quick fix. The rhetorical benefits of formal assessment are to articulate standards, focus instruction, motivate students and staff, hold feet to the fire, etc. The feared costs of such assessments include reducing to trivia the important goals of education, increasing the dropout rate, generating systematic attempts to "get around" the mandate, narrowing the curriculum, and so on. Studies of acutal effects of testing reforms will be released shortly and some light may be shed on the utility of assessment as a productive instrument of educational change.

Assessment as a Quality Indicator

The use of student achievement is a legitimate important indicator of educational quality. If they are to be used as part of a system of higher education, student assessment programs must be constantly held to their purpose: to provide an accurate and representative reflection of educational quality. Methodology used in student assessment does not meet this purpose. In my view, student assessment programs must intrinsically relate to real instructional programs in departments and courses. They must reflect the diversity of our offerings and what students learn from their coursework and their college experience. At present, we have relatively little evidence to document the effects of our educational efforts in higher education. I believe we can collect such evidence in a way that will avoid the bureaucratic and irrelevant character of much top-down assessment. We should try to avoid the use of omnibus assessment, where a single instrument is purported to be a major valid indicator of quality. The nature of higher education is such that using a single common measure to reflect student learning will provide very little valid information about educational quality. Most everything will be missed. We may, better still, find a way to use student performance assessment as a powerful instrument of improvement.

Developing an Approach to Individual Instructional Assessment

The model for student assessment in higher education I propose is one that incorporates student assessment as part of the teaching mission of the institution (Cross, 1986). Its purpose is to contribute to the development of educational quality. Call it individual instructional assessment (IIA). IIA develops from a view that colleges and universities have teaching responsibilities to individual students. The teaching responsibilities for individual students get executed as students relate to one another, to professors, to teaching assistants, and to other institutional resources. The product of this individual experience is what we should assess. Even though teaching is sometimes a mass act, its reality occurs in the complex interaction among the students and all these resources (Pace, 1985). To acknowledge and assess the individual, distinct, personalized nature of this experience is critical. However, such acknowledgement should not be confused with models of instruction (such as those advanced and tested by Keller (1969) and Bloom (1967; 1984). IIA does not presuppose self-paced instruction and is independent of instructional strategy. The purpose of IIA is to use assessment as a way to recognize and extend individual student accomplishment rather than to homogenize it. Its slogan was promulgated by Judah Schwartz (1978), in other contexts, some years ago: "People come in groups of one." So do higher education institutions.

A new approach to student assessment in postsecondary education is needed. This approach would use as its centerpiece the specific accomplishments of students in academic courses and courses of study, instead of their performance on specially constructed, mandated measures. So I will not discuss today a procedure to develop particular instruments. Outcomes of higher education would be documented by providing a wide range of *examples* of the kind of work accomplished by students at various levels and majors. The system would not be uniformly applied to all courses, nor would exhaustive reporting be expected. Rather an *institutional portfolio* would be created. If numbers are required, as they almost always are, frequencies of students performing at the illustrated level or above would be provided for the academic majors assessed. It is bottom-up demonstration of quality, clearly superior, I think, to judgments made on the basis of transcript analyses or catalog review.

The characteristics desired of such measures are obvious. The common, casually developed tests of knowledge and information in rampant use could realistically provide only a piece of the information. New, carefully

developed tasks for essay examination or term papers would be prepared. Criteria for judging the quality of responses would also be articulated. In operation, these assessments would be administered on a schedule naturally demanded by course organization. Feedback to students would be provided rapidly and in a way that strengthens the personal nature of the college experience.

What those tasks should be and the form of the feedback should be a faculty matter. Educational quality, in terms of what and how well students learn the full range of academic offerings will thus be directly affected. As present institutionally-generated student assessment is focused on scheduled, quantitative summaries of students' performance, IIA is periodic, qualitative, formative, diagnostic, and informative. IIA would also serve to increase rather than decrease the range of approaches used to assess learning. It also has particular strength as a means to provide careful differentiated feedback for students.

Of course, such a position requires a massive effort to train faculty members. They need to see that the way they assess students communicates what they view as important to learn. They need to believe that careful, timely, and personalized feedback can transform the college experience for students. They need to see assessment as more than a means to grade students or to meet bureaucratic requirements. It must contribute to their teaching effectiveness.

Do faculty care enough to engage in the serious work of developing high quality measures of course performance? We know they are relatively unskilled now. Whether some would embrace the use of high quality measurement approaches (such as domain-referenced assessment) remains to be seen.

What conditions are required for such a system to work?

- Agreement from top management that such an approach would directly rather than indirectly both impact and reflect higher education quality and that it is worth doing and superior to approaches using single measures.
- Incentives for faculty to take this responsibility seriously.
- A plan for institutional development, first to find leading academic institutions willing to undertake a pilot effort, and, within institutions, prestigious academic departments to provide the model for others.
- Useful approaches, tools, and training procedures from the measurement community.

Necessary Contributions from the Measurement Community

Colleges and universities, if they were to take seriously and systematically the charge to improve educational quality, need certain assistance from the measurement community. For example, approaches to the measurement of deep understanding of subject matter would need expansion. In a project in this domain we are attempting to develop procedures for assessing essays and term papers that incorporate appropriate cognitive representation of subject matter (Baker & Herman, 1986), reliable, and valid scoring of student responses, and procedures that do not demand inordinate time to evaluate each student's effort (Quellmalz, 1984). The measurement community needs to expand the options it offers college professors to assess subject matter and cognitive understanding.

Secondly, technological supports to the development of assessments are at least on the drawing board (Baker & Linn, 1985). The search should intensify for procedures to use computer technology to represent subject matter knowledge and to develop locally appropriate measures of student performance. As part of new CRESST Centre for Research on Testing, we have a design project to explore techniques from artificial intelligence to create a test developer assistant (Baker, 1986).

Third, help from offices of institutional research and evaluation is needed to provide the structure and training required for such an experiment to work.

Summary of Potential Effects

If successful, the results of IIA should be:

- to deepen the sense of intellectual engagement of students by requiring of them high level, defensible performance, and by providing timely individualized feedback,
- to stimulate faculty reflection on the real teaching mission of colleges and universities,
- to avoid the use of marginally valid measures in the assessment of higher education, and
- to provide appropriate indicators of higher education quality, in the form of institutional portfolios.

In this way, we can contribute to the responsible assessment of our higher education institutions. We must recognize that our institutions are complex, our students are different, and that our assessment approaches need to reflect those complexities.

References

Baker, Eva L. "The Impact of Advances in Artificial Intelligence on Test Development." *Continuation Proposal for OERI Center for Student Testing, Evaluation and Standards*, Los Angeles, CA: UCLA, 1986, pp. 82-87.

Baker, Eva L., & Joan Herman. "Issues in Content Assessment." *Continuation Proposal for OERI Center for Student Testing, Evaluation and Standards*, Los Angeles, CA: UCLA, 1986, pp. 74-81.

Baker, Eva L., & Robert L. Linn. *New Testing Technologies*. Los Angeles, CA: Office of Technology Assessment Study on Standardized Testing, UCLA, 1986.

Bloom, Benjamin S. "The Search for Methods of Group Instruction as Effective as One-on-One Tutoring." *Educational Leadership*, May 1984.

Gross, Patricia K. "A Proposal to Improve Teaching." *American Association for Higher Education Bulletin*, Vol. 39, no. 1, Sept. 1986.

Keller, F.S. "Goodbye Teacher." *Journal of Applied Behavior Analysis*, 1968, 1, 79-89.

Pace, C. Robert. "Perspectives and Problems in Student Outcomes Research." *Assessing Educational Outcomes*, 47, 1985.

Quellmalz, E. "Designing Writing Assessments: Balancing Fairness, Utility and Cost." *Educational Evaluation and Policy Analysis*, 6(1), 1984.

Schwartz, Judah L. "Assessment that Respects Complexity in Individuals and Programs." Paper prepared for the National Conference on Urban Education, held in St. Louis, Missouri, July 10-14, 1978.

U.S. Department of Education. *Indicators of Education Status and Funds*, Washington, DC: P. A-6, 1985.

The Case for Unobtrusive Measures

PATRICK T. TERENZINI
University of Georgia

There can be little doubt that much of what we know from the social sciences has been developed from interview or questionnaire data. But what can we say about the fidelity of those portraits for the social and educational behaviors and phenomena they depict? Consider the following:

> ...research in intelligence testing show(s) that dependable gains in test-passing ability (can) be traced to experience with previous tests even where no knowledge of results (has) been provided...Similar gains have been shown in personal adjustment' scores (Webb et al., *1966*, p. *19*).

> Male interviewers obtain fewer responses than female, and fewest of all from males, while female interviewers obtain their highest responses from men, except for young women talking to young men (Benney, Riesman, & Starr, *1956*, p. *143*).

> Sequences of questions asked in very similar format produce stereotyped responses, such as a tendency to endorse the righthand or the lefthand response, or to alternate in some simple fashion. Furthermore, decreasing attention produces reliable biases from the order of item presentation (Webb et al., *1966*, p.20).

Thus, much of what we know may be biased in various and sometimes unknown ways. But if what one blind man learns about elephants is biased by the data-gathering procedures adopted, measurement and sampling theory suggest it is reasonable to expect that the evidence gathered by multiple blind men, when pooled, will give a better, if imperfect, approximation of an elephant. There is, after all, more than one way of knowing. The central thesis of this paper is that multiple research designs and measures of educational outcomes are more likely to yield reliable and valid assessments of educational outcomes than is the current reliance on

interviews and questionnaires.

Consider the following:

The wear in the floor tiles in Chicago's Museum of Science and Industry.
The shrinking diameter of a circle of seated children.
Pupil dilation in the eyes of jade customers.
The bullfighter's beard.

Each of these conditions can, under certain circumstances, be taken as a measure of a phenomenon of interest to someone. Taken from Webb et al. (1966), each is an example of what has come to be called "unobtrusive measures," a general class of measurements presumed to reduce or eliminate the potential for reactive bias: responses uncharacteristic of the attitude or behavior outside the measurement situation and induced by the measurement act itself. The premise is that when interviews and questionnaires are used in social science research, the process of data collection intrudes itself into the consciousness of the subject and, as a consequence, alters the subject's responses. Unobtrusive measures, by their nature, avoid most, if not all, of the reactive bias associated with interview and questionnaire methodologies.

Webb et al. (1966) have described five categories of unobtrusive measures: physical traces (natural erosion or accretion processes, such as the wear on library book pages or the refuse left behind by an earlier civilization); continuous archival records (e.g., actuarial records, government records); intermittent archives (e.g., written documents, sales records); simple observations (e.g., of behaviors), and physical devices (e.g., cameras, video and audio tapes).

The measures listed above index some interesting illustrations of physical traces and simple observations. For example, the fact that the floor tiles around the hatching-chick exhibit require replacement approximately once every six weeks, compared to a replacement rate of several years for the tiles around other exhibits, can be taken as a reasonably clear reflection of the relative interest-value of the exhibit. So far as the shrinking diameter of the circle of children is concerned, if it were also known that the shrinkage was observed during a ghost-story-telling session, then the observation would have been recognized for what it is: an unobtrusive measure of the degree of fear induced in the children by the stories (and how much more reliable and valid than what the children might tell us if asked, "How scared were you?"). As for the dilation of the pupils in customers' eyes, Chinese jade dealers have used it as an indicator

of customer interest in various stores. And bullfighters' beards have been observed to be longer on days when the matador must enter the ring. There is no consensus whether the longer growth is attributable to higher anxiety or to whether he simply stands farther from the razor on those days. Probably both (Webb et al., *1966*, pp. v and 2).

Much has been written on how the methods of the social sciences might be brought to bear on questions of outcomes assessment in higher education (e.g., Ewell, *1985*, *1985*; American College Testing Program, *1980*; Astin, *1977*). Less attention, however, has been given to the measurement problems inherent in these methods and to how those problems might be avoided or at least counterbalanced. Some critics consider the present reliance on interviews and questionnaires to be both unwise and unnecessary. Webb et al. (*1963*), for example:

> lament this overdependence upon a single, fallible method. Interviews and questionnaires intrude as a foreign element into the social setting they would describe, they create as well as measure attitudes, they elicit typical roles and responses, they are limited to those who are accessible and will cooperate, and the responses obtained are produced in part by dimensions of individual difference irrelevant to the topic at hand.
>
> *But the principal objection is that they are used alone* (p. 1; emphasis in the original).

Unobtrusive measures, such as those listed above, offer an important methodological counterweight to the unknown and unbalanced reactive bias in interview- and questionnaire-generated data sets, such as those upon which we now rely to assess the educational outcomes of college.

The remedy for these ailments, of course, lies not in the replacement of the research tools now in widespread use. This is no call to rally the Assessment Luddites. Rather, the intention is to encourage outcomes researchers to supplement standard approaches with methods and measures now largely unknown, unconsidered, or ignored. The purpose, here, is to make "The Case for Unobtrusive Measures," and that warrant can be argued on at least three grounds (one major, and two secondary): 1) measurement, 2) cost, and 3) prudence.

The Measurement Warrant

The strongest arguments for the use of unobtrusive measures can be made

(appropriately enough) on measurement grounds. Recall that the principal objection of Webb et al. (1966) to the current reliance on interviews and questionnaires was that "they are used alone" (p. 1). The foundation of this objection is that:

> Every measurement procedure carries with it certain characteristic sources of error...it follows that they are in error in different ways and different degrees. The errors we refer to are constant *within* types of measures—the direction and size of the error are assumed to be fixed for a given set of measurement operations. However, the direction of errors is assumed to be random *across* procedures. For any given measurement task, the errors are additive: an error in one direction will tend to cancel out an error in the other direction (Sechrest and Phillips, 1970, p. 2).

Sechrest and Phillips go on to note problems occasioned by differences in the magnitudes of the errors involved and their effects on the precision of measurement, but the point is clear and the strongest argument for the use of multiple and *different* measures of the same trait or behavior—what Webb et al. (1966) and others (e.g., Campbell & Fiske, 1959) refer to as "multiple operationism." The intent is to employ multiple measures that "share in the theoretically relevant components (of the trait or behavior under study) but have different patterns of relevant components" (Webb et al., 1966, p. 3). When one samples measures, one also samples their strengths *and* their weaknesses. And as in sampling theory, the larger the sample size, the greater the reliability of estimation.

The utility of multiple measures in general, and unobtrusive measures in particular, is apparent in another way. Much of the research on student outcomes, particularly that focusing on institutional contributions to student growth, relies on various causal modelling techniques based on multiple regression. The multicolinearity among theoretically independent predictor variables, and the autocorrelations among the same measures used over time in longitudinal designs, present well-known, but frequently ignored, problems for the interpretation of path coefficients or regression weights. The problems of "bouncing betas" and the difficulty of replicating most studies in the social sciences are also well-known. Such interpretive difficulties notwithstanding, however, one researcher (cited in Kerlinger & Pedhazur, 1970, p. 446) has suggested that regression coefficients give us the laws of science, and many who employ regression analysis, or who read and rely on the results of such studies, may be similarly inclined to place more credence in the findings than is warranted.

The wisdom of multiple—and unobtrusive—measures is evident in still other ways. Research on the dynamics of attitude and value formation and change has both perceptual and behavioral dimensions. What correspondence exists between what a respondent professes to believe and how that person actually behaves? Reliance on questionnaires and interviews in such investigations requires an act of faith that the correspondence is high, when the fact of the matter may very well be otherwise. One can have significantly greater confidence in the reliability and validity of interview- or questionnaire-based claims about attitudes and beliefs if those claims are manifested behaviorally in natural settings. Used in this fashion, unobtrusive measures constitute a form of convergent validation and go a long way toward reducing the internal validity problems inherent in *ex post facto* research designs.

Unobtrusive measures have their own limitations, of course, for we rarely, if ever, know *their* characteristic sources of error. Thus, we cannot confidently estimate the extent to which use of an unobtrusive measure would be a useful and complementary addition to a series of measurement procedures or simply increase the error already present. And, like interview and questionnaire items, to the extent that unobtrusive measures rely on single observations, they are likely to be unreliable and, consequently, of limited validity (Sechrest & Phillips, 1979, pp. 5-7). Despite more than a two-decade history, much research remains to be done on the measurement characteristics of unobtrusive measures.

Before all hope and confidence in the utility of unobtrusive measures is abandoned, however, it is useful, at least insofar as the assessment of educational outcomes is concerned, to differentiate "unobtrusive measures" as a set of scientific research tools from "unobtrusive measures" as a metaphor. In the first instance, it is quite possible to apply unobtrusive techniques and measures in a remarkable variety of experimental studies (see Bochner, 1979). As such, the rigor characteristic of true experiments can be brought to bear in naturalistic settings (like colleges and universities) and threats to internal validity are significantly reduced if not eliminated.

For example, if an institution wished to know the extent to which cultural and racial openness was a trait characteristic of the campus, one might design a study similar to that reported by Campbell, Kruskal and Wallace (1966). In that investigation, the tendency of White and Black college students to sit by themselves in racially homogeneous groups in classrooms (rather than mixing randomly) was studied as an indicator of racial attitudes.

While such formal, unobtrusive research efforts are certainly possible, they are probably not likely to comprise a complete or adequate outcomes assessment program. "Unobtrusive measures" as a metaphor for non-reactive sources of information that *already exist* in various forms and locations across a campus are more likely to yield useful vehicles of assessment. Examples include such standard records as registrar's files, disciplinary records, Graduate Record Examination (GRE) scores, and alumni giving records. The category can also include less conventional measures, however, ranging from transcripts sent to other undergraduate institutions (student satisfaction), to case loads in the health services and psychological counseling service (amount of stress on campus), to library usage rates (students' intellectual curiosity). Unobtrusive measures may be based on observations as well as records. Such measures in colleges and universities might include assessment of a campus's intellectual climate as revealed on bulletin boards and in graffiti (see Ciardi, 1970, for a delightful discussion on this topic) and in conversations overheard in a student union snack bar. The point to be made is that unobtrusive measures—whether scientifically formal or casual—offer a source of information about the educational process and its outcomes that serves a legitimate and important measurement role by counterbalancing the systematic error characteristic of conventional measurement and research designs and by validating information gathered by means of those standard procedures.

The Cost Warrant

The costs of assessing educational outcomes are little understood. The proponents of the "benefits" portion of the cost-benefits equation have been dominant, and only recently has attention been turned to an estimation of the other side of the balance scale. How much in the way of resources is and should be invested in the production of outcomes information? The question applies to all information gathering, of course, whether outcomes or otherwise, but costs in other sectors are better understood and estimated than they are in outcomes assessment. The real issue, as Ewell and Jones (1986) put it, is: "How much *more* money (beyond that already committed to outcomes-related information gathering) do we have to spend to put in place an assessment program that is appropriate to our needs?" (p. 34).

Based on a set of assumptions about the nature of the assessment

programs likely to be mounted by institutions of varying types and sizes, Ewell and Jones (1986) estimate incremental costs ranging from $30,000 (in a small, private, liberal arts college) to $130,000 (in a major public research university). It is important to bear in mind that these are incremental, not total, cost estimates. It is revealing to notice Ewell and Jones' assumption of the use of conventional questionnaires, whether commercially available (e.g., The ACT's COMP) or locally developed (e.g., senior examinations in the major field disciplines).

No one has attempted to estimate the incremental costs of assembling information unobtrusively. Given the fact that much of this sort of data already exists, and given that much of it is electronically stored and retrievable, it seems reasonable to suggest that the costs of unobtrusive measurement and analysis are likely to be lower than those of more conventional measures and methods, perhaps significantly lower. There is, of course, considerable room for cost variability, but the initial proposition holds: analyzing data that are already available in one form and place or another is likely to be less costly and time-consuming than gathering data *de novo*.

The Prudence Warrant

Ewell (1984) has written that "the most vehement objections to the systematic assessment of institutional impact will come from faculty" (p.72). These objections, says Ewell, are likely to derive from either or both of two sources: first, the fear of being negatively evaluated, and second, a philosophical opposition based on the belief that the outcomes of college are inherently unmeasurable and that the evidence from such studies is "misleading, oversimplifying, or inaccurate" (p. 73).

To counter faculty opposition, Ewell recommends that persons responsible for outcomes assessments "recognize publicly the inadequacy of any *single* outcome measure or indicator and . . . collect as many measures of program effectiveness as possible" (p. 73). The point is related to the argument for unobtrusive measures made earlier on measurement grounds and is likely to be recognized and given weight by faculty of all disciplines. The effect is likely to be a reduction in faculty resistance to educational assessment. Even if the measurements cannot be easily explained to non-social scientists, most faculty members will be familiar with the concept of "triangulation" in astronomy, as well as in map-reading and surveying. The use of multiple measures to portray some educa-

tional outcome is likely to have a face validity that is appealing to faculty members. It seems reasonable to expect such an effect to influence positively both faculty participation in outcomes assessment programs and confidence in the conclusions derived from the evidence assembled.

Unobtrusive Measures in Higher Education

What are some unobtrusive measures in higher education and how might they enhance our understanding of various educational outcomes? Ewell (1984), following a review of various structures and taxonomies, has suggested that educational assessment should focus on three major areas: knowledge, skills, and values and attitudes, with a fourth category, students' relations with various groups in the larger society, representing the behavioral manifestations of the first three areas. Juxtaposition of these four dimensions against three of the general types of unobtrusive measures described earlier affords a useful framework for thinking about the sorts of institutional information that might be used to aid educational assessment. The matrix below is intended to be suggestive, to focus thinking on important assessment topics, and, thereby, to highlight the potential opportunities to employ unobtrusive measures.

Types of Unobtrusive Measures

Outcome Categories	*Physical Traces*	*Archives & Records*	*Observations*
Knowledge			
Skills			
Attitudes/Values			
Relations w/Society			

Space precludes discussion of possible measures that might occupy each of the cells in this matrix, and, as will be seen, the boundaries between the several categories of unobtrusive measures are not always precise. Moreover, some of the cells are of greater interest than others, and some unobtrusive measures are more readily accessible than others. Two cells easily meet both of these criteria, namely, the "Knowledge-Archives" cell and the "Attitudes/Values-Observations" conjunction, and attention will be focused on them for illustrative purposes, beginning with the latter of the two cells.

The observational techniques of Campbell, Kruskal and Wallace (1966) for inferring racial attitudes and relations on a campus have been summarized. Variations on this approach might include a study of "aggregating" (Campbell, Kruskan and Wallace, 1966) in dining halls and cafeterias, in clusters of students studying in the library or gathering in other public areas, in institutional residence hall roommate patterns, and in other institutional settings.

Something of the importance students attach to the life of the mind might be inferred from several sources, including the number, size, and participation rates in formal student organizations and clubs that have some specific, academic purpose (e.g., discipline-based clubs, literary and artistic publications, performing arts groups), as compared with organizations that have athletic, recreational, entertainment, social, or other purposes as their principale *raison d'être*. (Some of this information might be gleaned from records.)

Similarly, inferences about the relative emphasis given to the academic and social life of a campus might be made based on an examination of the content of campus concert, film, lecture, and speaker series, as well as attendance records. For residential campuses, the institution's role in students' lives—and its potential for influence—may be reflected in the extent to which students evacuate the campus for other locations on weekends. Ciardi (1970) has suggested that the content of graffiti reflect the intellectual tenor of a campus. One might add the content of bulletin boards to that reflection.

The number of students who are registered—and active—voters can be taken as a sign of students' interests in, sense of responsibility toward, and willingness to participate in the political life of a larger community. Or one might explore the level of social responsibility in a student body by designing an experiment around the frequency with which students returned library books that were presumably "lost." More simply, the proportion of the library's total overdue volumes that are signed out to students (or faculty) provides at least one index of the level of simple courtesy, if not social responsibility, on a campus. Vandalism, both in absolute magnitude and rate of change over time, offers another reflection of the quality of life and the attitudes and values prevalent on a campus. As suggested earlier, the rates over time at which the health service's physicians prescribe stress-related medicines, and variations in the case loads of the counseling center staff, might both be used to index the amount of potentially unhealthy—and perhaps educationally dysfunctional—stress in the campus environment. Hodgkinson and Thelin

(1971) offer an impressive list of other possibilities. The variety is limited only by one's imagination and ingenuity.

Without question, the major impact of college on students' cognitive development is delivered through the curriculum, and any outcomes assessment program must deal in one fashion or another with the curriculum and with classroom-based learning. A variety of reactive measures have been developed to assess the nature and extent of students' cognitive growth (e.g., the ACT-COMP and Graduate Record Examinations subject tests), and these measures are typically used in "value-added" research designs of varying degrees of sophistication and validity.

Warren (1984) and Pascarella (1986) discuss some of the conceptual and methodological limitations of this approach to educational assessment, and those critiques need not be reiterated here. The point to be made is that something of the nature and extent of student learning can also be inferred from unobtrusive measures, from a data base that already exists and that has reasonable claims to reliability, namely, the registrar's file, which contains extensive information on the courses students have taken and the grades received.

Fincher (1984), recognizing the weaknesses and disadvantages of the grade-point average as a criterion of what has been learned, also marvels that "it works as well as it does" (p. 380). He writes:

> . . .the freshman GPA will often display scalar features that are quite remarkable: a tenacious arithmetic mean, a standard deviation of about one-half letter grade, and a range of five or more standard deviation units. More remarkable, perhaps, the freshman GPA appears to be more immune to contamination than separate course grades are, and it is a relatively independent criteria despite being a faulty one. In addition, the freshman GPA is relevant to such educational decisions as the dean's list, student probation and dismissal, the maintenance of athletic eligibility, the continuance of scholarships, etc. If not a completely adequate criterion of academic performance, the freshman GPA still serves many educational purposes (p. 380).

Wilson (1983) reports that admissions measures are essentially as valid for predicting long-term GPA as freshman year GPA. Because of this property, Fincher suggests, cumulative GPA may yet be a useful measure in educational assessment and worthy of analysis. It might, for example, be used as the criterion in regression models and covariance analyses in which pre-college academic aptitudes and achievements (and other poten-

tially confounding variables) have been controlled in a study of the residual variance in long-term GPA attributable to student effort and to instruction and student learning. Similarly, if pre-college predictors of academic performance are found to have high multiple correlations with actual college achievement, reasonable suspicions might be raised about the overlap of high school and college coursework (Fincher, 1984).

The registrar's files offer other possibilities. For example, an examination of the distribution of courses taken by size and type of instruction (e.g., lecture, seminar, lab, independent study) might prove extremely revealing of the nature of the formal educational process experienced by students (e.g., graduating seniors). How many opportunities were there for students in small numbers to study with a faculty member? Such a review might focus on students' first two years. Do large lecture sections dominate students' early contacts with faculty and collegiate instruction? What is the relative balance of opportunities for active vs. passive student participation in their own learning? While recognizing that "small" is not necessarily "better," most faculty and administrators would probably be concerned if students' opportunities for small-group instruction were rare.

Examination of the relative proportional distributions of students majoring *and graduating* in particular disciplines will tell something of the nature of the educational program being delivered, and comparisons of such distributions, both one with the other and each over time, will detect shifting emphases in what students are interested in and what the institution is providing. Similarly, student retention rates, both within and across majors, may yield useful information. While such rates must be interpreted with considerable care, rates occupying one or the other tail in the distribution suggest something about students' views of the education afforded in those programs. Precisely what an extremely low retention rate means may be open to dispute, but at the very least it calls attention to the need for further investigation.

Transcript analysis affords a more detailed examination of curricula structures and student course-taking patterns and brings one still closer to the substance of students' formal education. Using this technique, Blackburn et al. (1976) undertook a national study of changes in degree requirements between 1967 and 1974, exploring the amount, structure, and content of general education, and the structure and flexibility in selected major degree programs. They found, for example, that the typical baccalaureate degree recipient in 1974, compared to 1967, had taken about 22 per cent less coursework in general education.

Galambos et al. (1985), in a study of teacher education in the states

comprising the Southern Regional Education Board, used transcript analysis to compare the course-taking patterns of teacher education and arts and sciences degree graduates. They found that, on the average, teacher education graduates took proportionally fewer general education credits in all areas except the social sciences than did arts and sciences graduates. Their analyses also led them to conclude that "Given latitude, some students will ferret out the routes of least resistance to meet their general education requirements, and then pass the word on to others" (Galambos et al., 1985, p. 78). Such a finding on an individual campus is likely to be justifiable cause for a detailed—and important—review of general education courses and requirements.

The State University of New York at Albany used transcript analysis to test a belief prevalent among faculty and administrators that students were not gaining a "general education" because the only degree requirements were those of the major program; all other degree credits were electives. The analysis provided information of the average number and percentage of course credits taken by graduates of each academic department in each of some 20 content areas. Results indicated that, while students in certain major field areas were apparently avoiding certain content areas (e.g., natural and physical sciences, or foreign languages), the deviations from what most academicians might consider a "general education" were by no means so great as had been anticipated.

A variation on this approach is afforded by the following matrix (adapted from Blackburn et al., 1976, who also offer some useful classification rules):

	Per Cent of Courses Taken	
Type of Course	*Breadth*	*Depth*
Required		
Restricted		
Elective		
Full Elective		

Using this matrix, a computer-based analysis of the transcripts of all students, or of sub-groups of students (e.g., selected majors, transfers, freshmen), would afford several kinds of information. It would reveal something of the variety and depth of the course work to which students have been exposed during any given period of time in their college careers. In addition, it would suggest the relative control over students'

course-taking exercised by the institution and the major department, and one might expect considerable variation across departments even within the same institution. Blackburn et al. (1976) offer a useful variant of the above matrix that differentiates general education requirements and courses from those of the major field. If still another variation were adopted to take into account *when* the course work was taken (i.e., a matrix that has the same breadth and depth columns, but has for its rows the time dimension of course-taking, say, lower and upper division), information would be gained on whether students are taking "breadth" courses prior to the selection of a major, or later in their college years, perhaps *after* the major program requirements have already been satisfied. The timing issue is important to the educational purposes of "general education" requirements. Do the requirements exist to ensure that students have a broad exposure to the various disciplines and on the basis of which they can make a more informed selection of a major program? Or are the requirements intended primarily to ensure that students are exposed to a broad intellectual experience at *some* point before they graduate?

Warren (1984) has suggested the analysis might be taken a step further. One might be inclined to believe, for example, that such course-taking pattern analyses do not provide a sufficiently detailed portrait of students' academic experience, for such analyses tell nothing of what students have learned. Warren suggests that a reasonable approximation of what has been learned might be obtained by reviewing examination questions and major paper assignments in courses that recur in the pattern of requirements for general education or for a specific major field—whether those courses are elective or required. As Warren (1984, p. 13) notes: "No pre-enrollment, normative, or comparative information need complement it. The assertion is simply that Program X as typically completed by a known number of students produces the described learning." A certain amount of faith is required, of course—faith that examinations and paper assignments reflect course content and that a passing grade reflects the occurrence of learning above some threshold of acceptability.

It should be evident by now that researchers in higher education have a wide variety of research designs and measures upon which to draw in their efforts to assess the outcomes of a college education. Thus far, however, the record indicates a virtually exclusive reliance on a subset of those designs and methods. The purpose of this paper has been to suggest ways in which conventional methods of assembling information on student growth might be supplemented in ways that illuminate rather

than obscure. Webb et al. (1966, p. 34) put it succinctly:

> So long as we maintain, as social scientists, an approach to comparisons that considers compensating error and converging corroboration from individually contaminated (measures), there is no cause for concern. It is only when we naively place faith in a single measure that the massive problems of social science research vitiate the validity of our comparisons.

References

American College Testing Program. *College Outcomes Measurement Project (COMP): Summary Report of Research and Development 1976-80.* Iowa City, Iowa: American College Testing Program, 1980.

Astin, A. W. *Four Critical Years.* San Francisco: Jossey-Bass 1977.

Benney, M., D. Riesman, & S. Star. "Age and Sex in the Interview." *American Journal of Sociology, 1956, 62,* 143-152.

Blackburn, R., E. Armstrong, C. Conrad, J. Didham, & I. McKune. *Changing Practices in Undergraduate Education.* Berkeley: Carnegie Council on Policy Studies in Higher Education, 1970.

Bochner, S. "Designing Unobtrusive Field Experiments in Social Psychology." In L. Sechrest (ed.), *Unobtrusive Measurement Today, New Directions for Methodology of Behavioral Science, No. 1.* San Francisco: Jossey-Bass, 1979.

Campbell, D. T., & D.W. Fiske. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin, 1959, 56,* 81-105.

Campbell, D. T., W.H. Kruskal., & W. P. Wallace. "Seating Aggregation as an Index of Attitude." *Sociometry, 1966, 29,* 1-15.

Ciardi, J. "Graffiti." *Saturday Review, 53,* May 16, 1970, 10ff.

Ewell, P. *Information on Student Outcomes: How to Get It and How to Use It.* Boulder, CO: National Center for Higher Education Management Systems, 1983.

Ewell, P. *The Self-Regarding Institution: Information for Excellence.* Boulder, CO: National Center for Higher Education Management Systems, 1984.

Ewell, P. (ed.). *Assessing Educational Outcomes, New Directions for Institutional Research, No. 47.* San Francisco: Jossey-Bass, 1985.

Ewell, P., & D. Jones. "The Costs of Assessment." In C. Adelman (ed.), *Assessment in American Higher Education: Issues and Contexts*. Washington: U.S. Office of Education, Office of Educational Research and Improvement, 1986. (U.S. Government Printing Office. Document No. OR 86-301).

Fincher, C., "Educational Quality and Measured Outcomes." *Research in Higher Education*, 1984, 20, 379-382.

Galambos, E. C., L.M. Cornett, & H.D. Spitler. *An Analysis of Transcripts of Teachers and Arts and Sciences Graduates*. Atlanta: Southern Regional Education Board, 1985.

Hodgkinson, H., & J. Thelin. *Survey of the Applications and Uses of Unobtrusive Measures in Fields of Social Science*. Berkeley: University of California at Berkeley, Center for Research and Development in Higher Education.

Kerlinger, F.N., & E.J. Pedhazur. *Multiple Regression in Behavioral Research*. New York: Holt, Rinehart and Winston, 1973.

Pascarella, E.T. *Are Value-Added Analyses Valuable?* Paper presented to the Educational Testing Service Invitational Conference on "Assessing the Outcomes of Higher Education." New York, October 15, 1986.

Sechrest, L., & M. Phillips. "Unobtrusive Measures: An Overview." In L. Sechrest (ed.), *Unobtrusive Measurement Today. New Directions for Methodology of Behavioral Science, No. 1*. San Francisco: Jossey-Bass, 1979.

Warren, J. "The Blind Alley of Value Added." AAHE *Bulletin*. September 1984, 10-13.

Webb, E.J., D.I. Campbell, R.D. Schwartz, & L. Sechrest. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally, 1966.

Wilson, K.M. *A Review of Research on the Prediction of Academic Performance after the Freshman Year*. College Board Report No. 83-2. New York: The College Board, 1983.

Using Assessment to Improve Instruction

K. Patricia Cross
Harvard Graduate School of Education

According to the latest *Campus Trends* report issued by the American Council on Education (El-Khawas, 1986), three fourths of all college administrators think that assessment is a good idea whose time has come. That's interesting, but even more interesting is the finding that almost all college administrators (91 percent) think that assessment should be linked to instructional improvement. Most authorities on the subject of assessment share that conviction. Turnbull (1985, p.25) observes that "the over-riding purpose of gathering data is to provide a basis for improving instruction, rather than keeping score or allocating blame." And the report just issued by the Education Commission of the States (1986, p.32) asserts that "Assessment should not be an end in itself. Rather it should be an integral part of an institution's strategy to improve teaching and learning. . . ."

In the jargon of the trade, the call for formative evaluation is loud and clear. Ironically, practically all of the proposals and practices in assessment today involve summative evaluation. We hear a lot about how institutional assessment and statewide testing will show us what is wrong and make educators more accountable, but there are few proposals for formative evaluation to show us how to improve education in process. The report just issued by the National Governors' Association, for example, is entitled *Time for Results*, and it is a call for summative, bottom-line accountability. While formulative evaluation gets the praise, summative evaluation gets the votes.

If we are to use assessment to improve the quality of education, perhaps the most important question for me to address is what decisions should be made in order to improve instruction. Stated that way, it's not a question that most college administrators are ready to grapple with because instruction generally takes place in the classroom, and the

Appreciation is expressed to Harvard University and to NCRIPTAL, University of Michigan, for funds to help support work on classroom research.

classroom is considered holy ground in academe. One group of optimistic assessors observe that, "Assessment seems to be loitering expectantly in the corridors of higher education, thereby reinforcing the hope that it will soon enter the classroom to serve the learner" (Loacker, Cromwell & O'Brien, 1986, p. 47). At the moment, I don't see any signs that anyone is ready to fling open the classroom door and invite the assessors in. In fact, I suspect that one reason for today's high interest in institutional assessment is that it is one way of demanding attention to the quality of student learning without actually entering the classroom. We in higher education have been especially reluctant to address the classroom performance of teachers for a number of reasons.

In the first place, we equate academic freedom with the sanctity of the classroom, and there is a tradition of restraint in probing too deeply what goes on there. Moreover, college teachers are authorities in their specialties. No one else in the institution knows quite as much about their particular subject as they do, and there is an understandable reluctance to tell faculty what or how to teach. And finally, there are some age-old questions that have not been answered to the satisfaction of many. What constitutes effective teaching? Who should evaluate college teachers and how? Can good teaching be recognized and appropriately rewarded?

If a major purpose of assessment is to improve instruction, can we use the results of assessment to do that? It doesn't seem very likely that we are going to reward individual teachers on the extent to which they demonstrate that they can "teach to the test," thereby pegging teachers' salaries to the scores of their students on assessment measures. Most people, I think, assume that assessment will improve instruction by documenting the strengths and weaknesses of student performance. Teachers will then use the results of the institutional assessment to take appropriate action. In higher education, "taking appropriate action" usually means making collective decisions about *what* is taught, i.e., about the curriculum. It rarely means doing anything about *how* it is taught. But *how* students are taught lies at the heart of quality education. It makes the difference between a lifelong learner and a grade grubber, between enthusiasm for learning and indifference to it, between an educated society and a credentialed one.

A few colleges, such as Alverno, with extensive experience and heavy faculty involvement in assessment, have managed to make a profound impact on teaching (see, for example, Loacker, et al, 1986), but most colleges, I predict, will conduct their assessment, add a few more course requirements, tighten academic standards, and see that students toe the

line. Assessment as currently conceived will probably stop short of the classroom door, doing little to improve the quality of instruction in the average classroom.

It is for this reason that I think institutional assessment needs to be accompanied by classroom assessment if we are to achieve long-term improvement in higher education. I have proposed elsewhere (Cross, 1986 a,b) the development of a new set of skills and tools that I call "classroom research." Its purpose is to help college teachers evaluate the effectiveness of their own teaching. The idea is to get faculty members involved *as individuals* in getting feedback from students on what they are learning in *that* classroom during *that* semester. Institutional assessment, in contrast, provides feedback on student learning college-wide, over a period of years, to faculty perceived as a team.

Ideally, a college is a community working in harmony toward common ends. Practically, it is a collection of individuals with maximum freedom to do their own thing, hopefully as well as they know how. The problem is that many college teachers really don't know how to teach very well. Typically, they have no training for teaching, and they have no one to talk with about it. While most now get student evaluations at the end of the semester (Erickson, 1985), they don't find the ratings very helpful in making changes in teaching methods (Clift and Imrie, 1980), and few have any skills for finding out what students are learning in their classrooms. Most are not even very proficient at getting maximum feedback on student learning from those two stalwarts of academe, final exams and term papers.

Thus, I contend that the most important decision that an institution can make regarding assessment is to explicitly move some of the decision-making into classrooms by giving teachers the necessary training and tools to assess what students are learning from them in the classroom.

Teacher involvement in classroom assessment is both necessary and desireable for a number of reasons.

First, it is by no means certain that a given teacher, say a professor of sociology, faced with results showing that students score below what might be expected on a test of social science will a) accept any personal responsibility for it or b) know what to do about it. One likely result of current efforts to measure "value-added" college-wide is to urge everyone to "teach to the test." That's not bad if the test truly measures the teaching aspirations of the college, but the better the test, i.e., the more it measures student growth and development, the more important teaching skill becomes. Many professors may discover that they don't know how

to "teach to a test" of personal or cognitive development.

One of the better ways to develop teaching skill is to provide feedback on what students are learning as a result of a given teacher's efforts. Realistically, the only way that can be done in higher education is to make the teacher responsible for formulating his or her teaching goals and assessing the results.

My second reason for encouraging classroom assessment is to add to our knowledge about teaching and its relationship to learning. As a profession, we don't know much about how to improve instruction. We struggle with faculty development programs and with disseminating the findings from research on teaching effectiveness, but most faculty teach as they were taught, and we're not sure how, or whether, to help them do differently.

There is a great deal of research on student evaluations of teaching and on teaching effectiveness. We know, for example, that college students are pretty good evaluators of teaching. They tend to give high marks to teachers from whom they learn the most (Centra, 1977; Cohen, 1982), they are reasonably unbiased, consistent over time, and in agreement with each other and with faculty evaluators (Gaff & Wilson, 1971).

Measures other than student evaluations also show agreement on the identification of effective college teachers. By this time, there has been enough research on teacher effectiveness that we can say with confidence that good teachers know their subject and their students. They are concerned about students, well-prepared, lucid, enthusiastic, available, and able to stimulate student interest and encourage their involvement in the work of the class (Abrami, 1985; Feldman, 1976; Kulik & McKeachie, 1975). Those are the results of literally hundreds of studies, and credible as they are, they are not very helpful to teachers. Even researchers who are presumably familiar with the research find it difficult to use the findings to improve their own teaching, and I know of no evidence that suggests that educational researchers are better teachers than those less well informed about research. While practitioners have been blamed for their failure to apply research, and researchers are regularly taken to task for failing to study questions that are relevant to teachers, the gap between research and practice is the fault of neither.

Educational research, with its search for general truths that hold across all classrooms, is not designed to address the situation-specific questions that teachers have. What a teacher wants to know is how his or her behavior affects the learning of a known group of students, studying a specific learning topic, under known conditions. But most research is

designed to control or eliminate those elements that pertain to a specific situation, and few researchers can afford to produce custom-designed research. By and large, the purpose of research in the social sciences is to push back the frontiers of knowledge and to build the foundations for understanding.

John Dewey (1929, p.19) wrote almost 60 years ago that "no conclusion of scientific research can be converted into an immediate rule of educational art." Research on teaching and learning is simply too large and complex to extract findings that can be easily disseminated to teachers as rules to improve teaching practices (Fenstermacher, 1982).

Donald Schon (1983) contends, in his new and provocative little book entitled *The Reflective Practitioner*, that research has done little to improve practice in any of the professions. In fact, he says, universities pursue "a view of knowledge that fosters selective inattention to practical competence and professional artistry" (p. vii). He calls for us to put aside the notion that "intelligent practice is an *application* of knowledge to instrumental decisions" (p.50) and instead to help professionals gain insight into their practice through an ongoing process of reflecting on what they know and articulating their intuitive thinking.

While it seems to me that Schon's reflection-in-action offers useful new perspectives on research to improve practice, I think it is both possible and desirable for teachers to collect and use both "hard" and "soft" data on student learning. Assessment designed for the improvement of teaching should be situation-specific, and it should provide immediate and useful feedback on what students are learning.

Situation-specific research may, at first blush, appear to result in knowledge with extremely limited usefulness to the profession of teaching, but my guess is that the exchange of knowledge from many specific classrooms will give teachers more useful insight into the teaching/learning process than the search for generalization across a "representative sample" of students, teachers, and subject matters. In any event, I think it highly likely that the knowledge gained from *doing* research is more likely to be used than that gained from *reading about* research.

My third and final reason for thinking that classroom assessment should be built into assessment programs, is to improve faculty morale through intellectual stimulation that is relevant to teaching. Unfortunately, the current lull in faculty hiring has convinced some institutions, historically committed to excellent teaching, that they should boost their academic prestige by hiring research faculties. The more likely result is that, as a society, we will sacrifice good teaching colleges for mediocre

research universities.

But so-called teaching institutions do have a problem in keeping a teaching faculty fresh and intellectually challenged. Heavy teaching loads tend to become repetitive, boring, and lacking in the intellectual stimulation that graduate students headed for careers in academe are taught to expect. Last fall's issue of *Change* magazine (September/October, 1985) presented a dismal picture of widespread demoralization of college teachers and pointed to what might be called the Rodney Dangerfield syndrome, "Teaching don't get no respect." If we are to make teaching institutions proud of their mission and to improve the status of teaching as a profession, we need to supply the tools for self-assessment and self-improvement, for those are the marks of a profession. Institutional assessment and state-wide assessment both carry implications of monitoring professional performance. Classroom assessment, carried out by teachers themselves, treats teachers as the professionals we want them to be.

In conclusion, one of the first rules of assessment, it seems to me, should be that the type of assessment information collected should be related to the type of decisions that it is possible to make. Since decisions about instruction are made by teachers, assessment should include information helpful in making decisions in the classroom. As a corollary, information should be collected as close to the source of potential action as possible. States can manipulate incentive systems and enforce standards. Institutions can see goals, establish climates, and reward behavior. Individual teachers, however, can relate teaching to learning, and that is the most important route to the improvement of undergraduate instruction.

References

Abrami, Philip C. "Dimensions of Effective College Instruction" *Review of Higher Education*, Spring 1985, 8 (3), 211-228.

Centra, John A. "Student Ratings of Instruction and Their Relationship to Student Learning." *American Educational Research Journal*, Winter, 1977, 14(1), 17-24.

Clift, J. C., and B.W. Imrie. "The Design of Evaluation for Learning." *Higher Education*, 1980, 9, 60-80.

Cohen, Peter. "A Validity of Student Ratings in Psychology Courses: A Research Synthesis." *Teaching of Psychology*, 1982, 9 (2).

Cross, K. Patricia. "Taking Teaching Seriously." Speech presented at the Annual Meeting of the American Association of Higher Education, Washington, D.C., March 11, 1986. (a)

Cross, K. Patricia. Speech prepared for the 1986 Conference of the Professional and Organizational Development Network in Higher Education. Hidden Valley, Pa., October 31, 1986. (b)

Dewey, John. *The Sources of a Science of Education.* New York: Liveright, 1929.

Education Commission of the States. "Transforming the State Role in Undergraduate Education." Denver, July 1986.

El-Khawas, Elaine. *Campus Trends, 1986.* Higher Education Panel Report No. 73. Washington, D.C.: American Council on Education, August 1986.

Erickson, Glenn R. "A Survey of Faculty Development Practices." Unpublished paper. University of Rhode Island, 1985.

Feldman, Kenneth A. "Grades and College Students' Evaluations of their Courses and Teachers." *Research in Higher Education,* 1976, 4, 69-111.

Fenstermacher, Gary D. "On Learning to Teach Effectively from Research on Teaching Effectiveness." *Journal of Classroom Interaction,* Vol. 17, No. 2 (1982) 7-12.

Kulik, J. A., and J. W. McKeachie. "The Evaluation of Teachers in Higher Education." In F. N. Kerlinger (Ed.), *Review of Research in Education,* Vol. 3. Itasca, Ill: F.E. Peacock, 1975.

Loaker, Georgine, Lucy Cromwell, and Kathleen O'Brien. "Assessment in Higher Education: To Serve the Learner." In *Assessment in American Higher Education.* Washington, D. C.: Office of Educational Research & Improvement, 1986.

National Governors' Association. *Time for Results: The Governors' 1991 Report on Education.* Washington, D.C.: August 1986.

Schon, Donald A. *The Reflective Practitioner.* New York: Basic Books, 1983.

Turnbull, William W. "Are They Learning Anything in College?" *Change,* November/December 1985, 23-26.

Are Value-Added Analyses Valuable?

ERNEST T. PASCARELLA
Professor of Urban Education Research, University of Illinois, Chicago

One of the watchwords of higher education in the late 1970s and on into the 1980s has been "accountability." Public institutions in particular are being called upon to document their effects in terms of student learning and development (Fincher, 1986; Hartle, 1986). Often the impetus for this call to accountability is state government and the higher education coordinating or monitoring boards which, in a number of cases, report directly to the governor or the state legislature. In my own state, Illinois, I have had the opportunity, for the past year, to be a member of a committee chosen by the Illinois Board of Higher Education to study the condition of undergraduate education in the state and make recommendations to the state legislature. Two pertinent paragraphs from the final draft of the report recommendations are as follows:

1. Colleges and universities shall conduct regular reviews of undergraduate education with emphasis on general education and the development of baccalaureate-level skills. The findings and conclusions of these reviews shall be reported to the Illinois Board of Higher Education.

2. Each college and university shall assess individual student progress in meeting the objectives of general education and the development of baccalaureate-level skills, and shall incorporate the results of assessment in the reviews of these areas. It is expected that colleges and universities will assess student progress at appropriate intervals and assure that assessment programs reinforce the maintenance of academic standards.

State-mandated student assessment in Illinois is certainly not an isolated case, as an increasing number of states are moving in a similar direction. Moreover, it is likely, based on the widespread impact of the recent National Institute of Education Report, *Involvement in Learning:*

Realizing the Potential of American Higher Education (Mortimer, et al., 1984), that the current emphasis on estimating the student outcomes of higher education will be with us for the forseeable future.

One of the more recognizable approaches to the assessment of student outcomes is the concept of "value-added." As I understand the term, value-added typically examines actual or inferred changes in students' performance over time. Students are assessed for entering competencies on some set of reliable and valid instruments and then are reassessed following a specified period of time (e.g., freshman to senior year) or the completion of certain courses, programs of study, or other educational experiences (McMillan, 1986). "Value-added" has another implication, however, which goes beyond simply looking at pre-post changes on some measure of interest. Specifically, it attempts to separate that portion of student growth or development which can be reasonably attributed to specific educational experiences from that attributable to confounding causes such as differential ability or simple maturation. In short, value-added entails an estimation of the "net effects" of college.

Of course this is my own understanding of what the term "value-added" implies. It must be acknowledged that its operational definition is likely to vary with who uses the term, in what context and for what purpose (Ewell, 1986). For some institutions, value-added may be inferred from simple freshman to senior changes in measures of cognitive development or learning (e.g., Lenning, Munday and Maxey, 1969; Dumont and Troelstrup, 1981). For other institutions, the value of the education they provide may be evidenced largely by the accomplishments and retrospective evaluations of graduates (e.g., Spaeth and Greeley, 1970; Pace, 1974). Still other institutions may see value added largely as I have defined it, and may concentrate on estimating the net effects of the collegiate experience (e.g., Mentkowski and Strait, 1983). Furthermore, even within the same institution, it can mean different things to different administrative, faculty, student and alumni constituencies. Who decides what is valuable, for whom it is valuable, and to what it is added?

Clearly value-added may not mean the same thing to all who use it. Indeed, as Ewell (1986) has pointed out, the operational definition of the term is still quite flexible and vague, and this vagueness may vary in direct proportion to its increasing use in public dialogue.

I. A Recent Debate on the Value of Value-Added Assessment

Perhaps in part because of the lack of a clear operational definition of the term, the concept of value-added has recently become the focal point of a spirited controversy. Some of the most respected names in the postsecondary scholarly community have chosen sides on the issue. Jonathan Warren (1984) has published an enthusiastic and cogent criticism of value-added. The gist of his argument is that, while in the abstract the logic of value-added has great appeal, in practice it seldom leads anywhere. The results of value-added analyses, he argues, "are often trivial, difficult to make sense of and peripheral to most instructional purposes" (Warren, 1984, p. 10). Warren's argument focuses largely on one level of analysis, the use of the value-added approach in assessing course-level instructional outcomes. Here he makes some telling points about the questionable practice of using pre-post differences as a measure of student learning in courses such as upper-division electromagnetic theory, where students can be assumed to have little pre-course content knowledge. Simply finding that students understand course concepts on the final examination is, he maintains, sufficient evidence to infer that most of the observed learning was a consequence of the course. A similar argument has been made by Pace (1985).

Warren's recommendation is that we need to abandon the unworkable concept of value-added and get on with alternative ways of assessing the effects of postsecondary education. Cameron Fincher (1985) is perhaps less convinced that value-added is a blind alley which needs to be abandoned, yet he is similarly skeptical in elucidating problems in its implementation. These problems, he argues, center on: 1) the development of reliable and valid instruments to measure various educational outcomes; 2) psychometric problems in assessing change; and 3) conceptual problems in interpreting college effects on achievement when most of the variation in achievement may be due to student aptitude, prior achievement and the quality of student learning effort rather than to instructional variables. While not necessarily ready to abandon the concept, Fincher is nonetheless skeptical about the ability of educators to apply value-added concepts to educational issues. He suggests that "if educators could agree on the assessable outcomes of higher education, take the time and effort to develop suitable forms of measurement and assessment, and restructure instructional efforts in terms of explicit instructional objectives, value-added concepts of education might then be

a solution to *some* educational problems" (Fincher, 1985, p. 398, author's emphasis).

Fincher's points are well taken. One must wonder, however, whether the issues which he elucidates are specific problems of a value-added approach to student assessment. A reasonable argument could be made that educational and behavioral research in general have traditionally been confronted with these and similar assessment issues. The knotty problems of instrument validity and attributing student learning to specific instructional practices are longstanding, if not adequately resolved, concerns of those interested in the effects of schooling at all levels (e.g., Wittrock, 1986).

Responding directly to Warren's (1984) article, Astin (1985b) and Ewell (1985) have defended value-added as an important contribution to our thinking about assessing the impact of post-secondary education. Both authors readily admit that the approach is not without its problems. At the same time, however, they argue that it has conceptual strengths which outweigh these problems. These include: 1) a focus on actual student development rather than typical measures of institutional "prestige" or "quality" (e.g., student body selectivity, resources per student, library size); 2) the requirement of clearly defining, in conceptual and operational terms, the desired outcomes of college or other educational experiences; and 3) systematic attempts at assessing the impact of educational experiences. They also argue, quite convincingly, that value-added as an approach to evaluating the impacts of college has been implemented in a systematic manner in only a few postsecondary institutions (e.g., Alverno College, University of Tennessee at Knoxville, Northeast Missouri State University). Thus, abandoning value-added now would be the equivalent of dismissing an idea with considerable rational appeal "before it has been more extensively tried, evaluated and is better understood" (Astin, 1985b, p. 11). Clearly, Astin and Ewell want to avoid throwing the baby out with the bathwater—at least until we have a chance to see how the baby matures.

Astin and Ewell make another point in their defense of value-added. Warren's critique, they suggest, is based almost exclusively on a single level of analysis, i.e., student learning in a single course. Ewell concedes the point that there may be course situations involving a well-defined, specialized body of knowledge to which students were not previously exposed where pretesting makes little or no sense. He further asserts that this is far from a valid critique of the entire value-added concept, however, since value-added is clearly useful in assessing cognitive and developmen-

tal outcomes of curricula and educational experiences more broadly conceived than individual courses. Indeed, it is in assessing these more broad-based effects that value-added has had its most typical application in postsecondary education.

It seems unlikely that the current debate concerning the utility of value-added assessment is over. The term itself is too value-laden and perhaps a bit overly pretentious in its public use by educators. In this sense it may appear to claim more than many current applications of value-added assessment have been able to deliver. As a result it will probably continue to invite spirited criticism which, in turn, adds fuel to an ongoing debate. If this debate is to contribute light as well as heat, however, it may be to our advantage to do two things. First, we might consider redefining value-added as the very fundamental and traditional research question: what are the student developmental outcomes associated with exposure to an educational experience which can be reasonably attributed to the educational experience itself and not to other factors? This is the "net effects" question, and it is far from being a new concern for educational researchers and evaluators (e.g., Campbell and Stanley, 1963; Feldman and Newcomb, 1969; Solmon and Taubman, 1973; Hyman, Wright and Reed, 1975; Bowen, 1977; Astin, 1977, 1982). Indeed, a basic issue in educational research and evaluation has long been the extent to which we can attribute student development to purposeful educational experiences. This is also, I believe, the core of what a value-added approach to student assessment is about.

Given this perspective, a second thing we might do is to place the discourse about value-added on a different level. If we are willing to accept value-added as a potentially important approach to the assessment of student outcomes, then it behooves us to consider ways in which the methodology of the approach might be enhanced and sharpened. The remainder of this paper will suggest and discuss a number of such methodological enhancements.

II. Pre- to Post-Changes: Improvement Can Be Misleading

It is often the case that the value-added or net effects of an educational experience will be inferred from pre- to post-changes (say from freshman to senior year) on some accepted measure of student development (e.g., critical thinking, moral reasoning, reflective judgment, ACT-COMP scores).

Unfortunately, even assuming the reasonable reliability of change scores, such mean changes reflect not only the effects of college, but also the effects of simple maturation. (Other variables such as history, instrument decay, and possibly even regression artifacts if the group is extremely low to begin with could also confound interpretation, but it is likely that maturation over time would be the major confounding variable.) This, of course, means that longitudinal freshman-to-senior changes probably overestimate the effect due to college alone, i.e., the unique effects of college (McMillan, 1986).

One possible way to deal with maturation is through the use of a control group of subjects followed over the same period of time, but not exposed to the particular educational experience (e.g., Plant, 1962; Telford & Plant, 1963; Trent & Medsker, 1968). In the situation where one is assessing institutional effects; however, reasonably comparable control groups not attending college are particularly difficult, if not impossible, to obtain. For such situations there are alternative cross-sectional or combined cross-sectional and longitudinal designs which provide reasonable controls for maturation. Consider the cross-sectional design where freshmen are compared with seniors on a measure of critical-thinking ability. The freshmen, who have not been exposed to the institution, would act as a control group for the seniors, who have theoretically benefited from four years of exposure to it. (To better reflect the entire college experience, the measure of critical thinking might be given to freshmen upon enrollment in the institution and to seniors in the final semester or quarter of their senior year.) The difference between the average freshman and average senior scores, statistically adjusted for differences in age, could be used to estimate the impact of the institution on critical thinking.

There are, of course, potential problems with this design that must be addressed. First, seniors probably represent a more selective group in terms of academic ability since a portion of the least academically proficient are likely to have flunked out or to have left for academic reasons. Second, there is the possibility of differential recruitment or admission criteria being used for the seniors versus the current freshmen (e.g., if the institution used a more stringent set of admissions criteria for the current seniors as versus the current freshmen, the former might be a more academically select group than the latter). This might also lead to systematic group biases on a factor such as academic aptitude, which, in turn, is likely to influence the level of critical thinking. While acknowledging that there is no ideal way to adjust for such pre-existing differences (Lord, 1967), one might nevertheless select the freshmen to be compared

from SAT or ACT ranges similar to those of persisting seniors, and accompany this with regression analysis to provide aptitude- as well as age-adjusted critical-thinking means for the freshman and senior groups. The difference between the aptitude- and age-adjusted means would likely yield a better, though still imperfect, estimate of net institutional effects, than that yielded by simply adjusting for age alone.

Cross-sectional designs such as this have recently been employed by a number of researchers in an attempt to separate college effects from those of maturation (e.g., Whitla, 1978; Mentkowski and Strait, 1983). It would also seem reasonable that such cross-sectional designs could be used in conjunction with pre-post longitudinal designs to provide a clearer picture of the influence of college versus the influence of maturation. Because the simple longitudinal results also include the possibility of maturational influences, they might be thought of as the upper-bounds or liberal estimate of the effect of college. Conversely, the adjusted cross-sectional results would tend to statistically remove any joint impacts of the college experience and normal student maturation; thus providing a conservative or lower-bounds estimate of the effect of college. The difference between the longitudinal and adjusted cross-sectional results might be used as an estimate of normal maturation during college. (In the absence of longitudinal data one might use the unadjusted differences between cross-sectional freshmen and senior cohorts to represent the upper-bounds estimate of college effects, the age and aptitude adjusted differences to represent the lower-bounds estimate, and the difference between the unadjusted and adjusted results to represent normal maturation.)

Another cross-sectional design which has been used to disaggregate college effects from age or maturation effects is one which takes advantage of the increasing numbers of older, nontraditional-age students in American postsecondary institutions. In this design, traditional-age freshmen (e.g., age 18), nontraditional-age freshmen (e.g., age 22), traditional-age seniors (e.g., age 22), and nontraditional-age seniors (e.g., age 26), might all be administered the measure of critical thinking. The effects of college versus maturation would be estimated by comparing the freshman to senior differences for both traditional and nontraditional students, with age differences between traditional and nontraditional-age freshmen and between traditional and nontraditional-age seniors. Examples of this design have been used in estimating the effects of formal education on the development of reflective judgment (Perry, 1970) by Strange (1978).

I offer these alternative designs not because they are without flaws,

Indeed, there are internal validity issues (i.e., what is the real cause of the effects observed) associated with both of them. As alternatives to simple pre- and post-assessments, however, they do provide somewhat greater control over the confounding influence of student maturation during college. Consequently, they probably provide more internally valid estimates of the value-added or net effects of educational experiences than do simple pre- to post-changes. In terms of contributing to our understanding of the value-added or net effects of educational experiences, pre- to post-changes in the absence of a control group are quite limited. One might conceive of them as a necessary but insufficient condition for documenting educational impacts. In most instances, the presence of a net educational impact will be accompanied by positive pre- to post-changes on measures of interest. One must be a bit cautious in this regard, however. Recent evidence reported by Wolfle (1983), for example, has suggested that the general effect of college attendance on mathematics achievement is to maintain precollege levels of competence. Net of other factors, those not attending college tend to decline in mathematic competence. Thus a college or educational impact may not always be accompanied by pre- to post-improvement, or even by pre- to post-change.

It is precisely in this type of situation that the term value-added can be misleading or even dysfunctional. As an estimate of college impact, value-added implies that something is added to the individual's level of development. In some areas of development, however, the impact of college (or other educational experiences) may be to prevent or retard decline rather than to induce major positive changes. Consequently approaches to value-added assessment which focus only on pre- to post-changes may be overlooking important college effects.

III. Direct and Indirect Impacts

I would argue that value-added assessment is made increasingly effective and useful in terms of its policy implications as it becomes more specific and focused. At the individual college level, this suggests the importance of identifying specific institutional experiences which enhance student development. Assessments for this purpose are often more fine-grained in conceptualization and analysis than those which are limited to determining the net effects of college or some particular educational experience. At the multi-institutional level, we have some exemplary applications of this

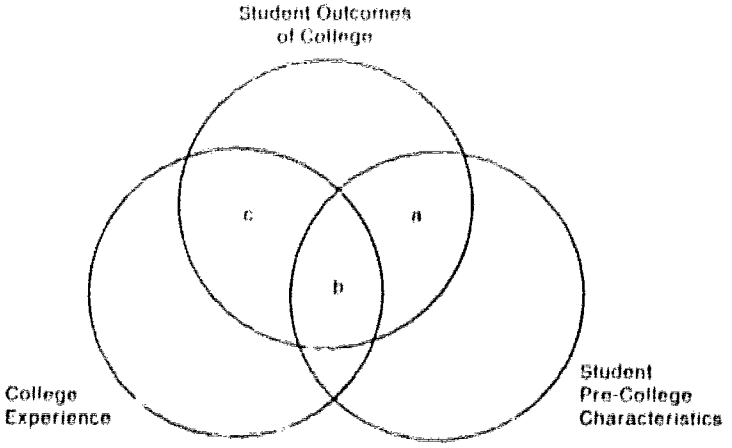
approach in the work of Astin (e.g., Astin, 1968, 1977, 1982; Astin and Panos, 1969) and Pace (1984). Generally this approach involves estimating the associations between various measures of the college experience (e.g., academic major, academic achievement, extracurricular involvement, interaction with faculty) and certain outcomes (e.g., standardized measures of achievement, graduate degree attainment, self-concept, values) with the confounding influence of student pre-college traits (e.g., aptitude, secondary school achievement, social origins) removed statistically. The result is an estimate of the effect of a particular collegiate experience independent of differences in student pre-college traits.

When multiple regression is employed to assess these partial associations between college experiences and college outcomes, the resultant regression coefficients (either standardized or metric) can be interpreted as estimates of the net or direct influence of a particular variable (Wolfle, 1985). Thus, in addition to estimating the extent to which an institution facilitates the development of critical-thinking ability from freshman to senior year, a value-added approach might also attempt to identify those particular institutional experiences that have nontrivial net associations with critical thinking. The results of these and similar analyses can provide useful information in terms of focusing attention on those potentially manipulable collegiate experiences that *may* causally influence the development of critical thinking (e.g., particular courses or course-taking patterns, interaction with faculty). It is important to stress the "may" in the previous sentence, since attributions of causality with correlational data are tenuous at best.

Such analyses are an important extension of value-added assessment and have been the analytic model for many benchmark studies of the influence of college on student development (see Feldman and Newcomb, 1969; Bowen, 1977). At the same time, however, these analyses are limited in their capacity to estimate the full range of a variable's influence. Because student pre-college characteristics and measures of the college experience typically have substantial correlations between and among themselves, there is usually a substantial portion of the explained variance which cannot be uniquely attributed to any particular independent variable. This is typically referred to as the joint or redundant effects of student pre-college traits and the collegiate experience (Cohen & Cohen, 1975). Figure 1 presents a schematic diagram showing the various unique (net) and joint or redundant effects.

Joint or redundant effects have often gone unanalyzed. In a classic paper, however, Alwin and Hauser (1975) suggest that when independent

Figure 1: Schematic of College Effects



a = Unique Effects of Student Pre-College Characteristics

b = Joint Effects of Student Pre-College Characteristics and the College Experience

c = Unique Effects of College (Independent of Student Pre-College Characteristics)

variables are in a causal sequence, unanalyzed joint effects may be attributable to effects transmitted through intervening causes. Recent evidence, for example, has suggested that exposure to preenrollment freshmen orientation may have little direct influence on first-year persistence/withdrawal behavior. However, such orientation experiences may facilitate initial student social integration in college which, in turn, positively influences persistence (Pascarella, Terenzini, & Wolfle, 1986). These are the indirect effects of a variable, and until comparatively recently they have been largely ignored in research on the impact of college on student development. With the increasing acceptance and use of causal modeling as a research and analytical methodology, however, this need not continue. Causal modeling, which is essentially an attempt to fit a theoretical, explanatory model to a matrix of correlations among variables, is the subject of a number of excellent discussions (e.g., Anderson & Evans, 1974; Heise, 1975; Wolfle, 1980; Maruyama & Walberg, 1982; and Wolfle, 1985).

Developing a causal model forces one to think theoretically, and, therefore, specifically and parsimoniously. One must specify not only the important variables (i.e., hypothesized causal influences) to be included in

the model, but also the causal ordering and the pattern of influences (causal paths) among variables. These relationships are then expressed as structural equations. Structural equations, which are typically solved by regression analysis, specify how each variable (including the criterion) is a function of causally antecedent variables in the model.

An important purpose of causal modeling is to portray the system of indirect as well as direct influences in a causal system. Consequently one is able to estimate not only the net direct influence of a variable on some outcome (i.e., the regression coefficient), but also the extent to which that variable influences intervening variables which, in turn, affect the outcome. The latter, of course, are the indirect effects, and are simply the sum of the products of direct effects through variables intervening between the variable in question and the outcome measure.

Because causal modeling permits one to portray the patterns of indirect as well as direct effects on some outcome, it yields a more complete estimation of the total effect (direct + indirect) of any particular variable. As such, causal modeling is a potentially important technique in value-added assessment, particularly if one is interested in understanding the process by which student development occurs rather than merely predicting its occurrence. It is frequently the case that, net of other influences, a particular variable may have only a trivial or non-significant direct effect on student development, yet its indirect influence may be substantial and statistically significant. By their very nature traditional regression analyses that focus on prediction will overlook this indirect influence, and lead to conclusions that the variable has an unimportant influence on student development.

Recent causal modeling analyses of the national Cooperative Institutional Research Project samples, for example, have suggested that college experience variables, such as place of residence and the size and complexity of the institution attended, have few if any net direct effects on student outcomes such as educational aspirations, academic self-concept, social self-concept, or humanitarian/civic involvement values. They do, however, have significant indirect effects that are transmitted through their influence on level of student social and extracurricular involvement during college. Similarly, while academic achievement during college did not directly affect the subsequent occupational or economic attainments of individuals, it did have an important indirect effect on them by enhancing educational attainment (Pascarella, 1985b, 1985c; Pascarella, Smart, Ethington and Nettles, 1980; Pascarella, Smart & Stoecker, 1986).

The point to be made here is not that value-added assessment should

be concerned with the particular variables in the above analyses. Rather it is to suggest that the concern of causal modeling with understanding the patterns of direct and indirect influences in a longitudinal process can provide a more complete and accurate estimation of the influence of specific educational experiences on student development. As such it is an important tool for sharpening the focus and increasing the understanding yielded by value-added analyses.

IV. An Example of the Use of Causal Models in Value-Added Assessment

One area where causal modeling might significantly enhance value-added assessment is in estimating the influence of differential coursework and curricular patterns on cognitive development. It seems reasonable to assume that the nature of one's academic program will be a major, if not the major, influence on learning and cognitive development during college. Recent evidence from secondary school samples, for example, suggests that differential quantitative course work accounts for much of the gender difference in Scholastic Aptitude Test mathematics scores (Pallas and Alexander, 1983). Thus, one might hypothesize that differential patterns of course work taken during college will have important direct effects on how much a student learns and in what areas he or she learns it (Pace, 1979).

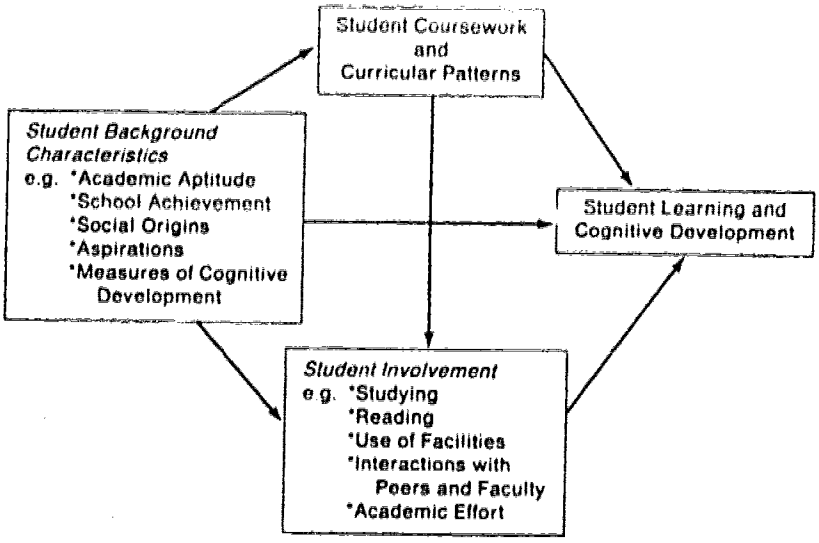
Beyond direct effects on learning, however, one might also be interested in how a student's academic experience indirectly influences cognitive outcomes by influencing the different dimensions of student involvement in college. Astin (1984) defines student involvement as the extent or amount of physical and psychological energy that the student devotes to the academic experience. Involvement includes not only time spent studying and in laboratories, but also time spent interacting with faculty and participation in cultural, artistic and extracurricular activities which enhance the intellectual impact of the institution. Much of the research of Astin (e.g., Astin, 1977, 1982, 1985a) and others (e.g., Pascarella, 1980; Terenzini & Pascarella, 1980; Volkwein, King & Terenzini, 1986) has underscored the importance of student involvement as an influence on cognitive development. Apparently a good deal of what students learn during college is the direct result of their own efforts. These efforts, of course, are no doubt a function of individual student attributes. Net of these attributes, however, they are also likely to be formed, in part at least,

by the nature of the student's academic and coursework experiences. Methods of organizing or quantifying student coursework or course-taking patterns are offered by Blackburn, Armstrong, Conrad, Didham & McKune (1976); Prather, Williams & Wadley (1976); Beeken (1982); McCombs & Smith (1986); and Ratcliff (1986).

Recently Pace (1984) has developed an instrument (The College Student Experiences Questionnaire) that measures the student's level of effort or involvement in various activities (e.g., studying, reading, attending cultural events, interacting with faculty and peers) during college. The CSEQ is essentially a series of scales that estimate the amount, scope and quality of effort students invest in using the salient facilities and opportunities provided by the institution. As such, the CSEQ is a potentially important instrument for assessing Astin's (1984) concept of "involvement."

Employing the CSEQ, one might develop an explanatory model of learning and cognitive development in college which posits that student coursework/curricular patterns are a function of student aptitude and various student background characteristics (e.g., social origins, educa-

Figure 2: A General Causal Model for Assessing the Effects of Student Coursework and Curricular Patterns on Student Learning and Cognitive Development During College



tional and occupational aspirations, gender). In turn, net of aptitude and background characteristics, coursework and curricular patterns would be expected to directly influence extent of student involvement. Finally, measures of learning and cognitive development at the end of college would be seen as a function of entering academic aptitude and student background characteristics, coursework and curricular patterns, and extent of student involvement. Figure 2 is a graphic portrayal of the general causal model.

Estimation of this model would permit determination of the total impact of differential coursework and curricular patterns on measures of student learning and cognitive development at the end of college. One would be able to estimate not only the net direct effects, but also the indirect influence through student involvement. The indirect effects would indicate the extent to which the institution's major structural mechanism for influencing student cognitive development (i.e., the academic program) does so by influencing students' involvement in their own learning. Such evidence might suggest ways of structuring student coursework or curricular patterns to maximize both the direct and indirect effects on learning and cognitive growth.

V. General Versus Conditional Effects

This issue concerns the level at which value-added or net educational effects are assessed. Most existing attempts at value-added assessment at the institutional level have assumed that the impacts of educational experiences are general, that is, that the impact is essentially the same for all students. This assumption certainly has the appeal of parsimony (i.e., other things being equal, the simplest explanation is often the optimal one). It can be argued, however, that assuming only general effects in one's analytic or assessment model ignores individual differences among students attending the same institution or exposed to the same educational experiences to produce conditional rather than general effects. Thus, the magnitude of the influence of certain educational experiences on student development may vary for students with different characteristics (e.g., level of entering aptitude, degree of prior exposure to, or competence in specific course content, level of intellectual orientation). Conditional relationships such as this might well be overlooked in assessment approaches which consider only general effects. In certain situations this may lead one to conclude that effects of specific educational experiences

are trivial, when in fact they may have pronounced positive effects for certain subgroups of students (Pascarella, 1985a).

The notion of conditional effects determined by the interaction of individual differences among students with different methods of teaching or the presentation of course content is a respected tradition in instructional research. Here it is typically referred to as aptitude x treatment interaction (Cronbach & Snow, 1977). Underlying its application in instructional research, however, is the more general perspective, supported by the psychology of individual differences, that not all individuals will benefit equally from the same educational experience. This idea may run somewhat counter to state-initiated mandates for accountability in which institutions are expected to demonstrate certain levels of effectiveness in promoting cognitive and other development for all students. Nevertheless, the consideration of conditional effects might well function to sharpen the focus of value-added assessment at the institutional level and enable it to better identify those particular students who are benefiting most or least from certain educational experiences. This information could then be used to focus institutional efforts on those student constituencies where its efforts appear to be least effective. Applications of the investigation of conditional effects in postsecondary education are shown in the work of Holland (1963) for career choice and academic achievement; Pfeifer (1976) for race and grades; Buenz and Merril (1968), Domino (1968), Pascarella (1978), Ross & Rakow (1981), and Stinard & Dolphin (1981) for different instructional approaches; Pascarella & Terenzini (1979) and Bean (1985), in research on student attrition from college; and Pascarella, Smart, Ethington & Nettles (1986) in research on the development of self-concept during college.

VI. Are Value-Added Analyses Valuable?

My answer to this original question of the paper is a cautious "yes." Undoubtedly, a value-added type approach can make potentially important contributions to our knowledge about institutional impact. My cautions have to do with the term itself and with the need for its rigorous application at the institutional level if we are to provide policy makers with accurate and useful assessments of institutional impact. As I have argued above, value-added is perhaps current and overly pretentious terminology for a long-standing and basic issue in educational research; namely, what are the net effects of educational experiences on student

cognitive and non-cognitive development? It may be to our advantage in the future to remind ourselves of this, for in doing so we can move more directly to substantive conceptual and methodological concerns.

Based on the above definition of the value-added approach, I have argued that a major concern should be on ways in which we can enhance the validity and usefulness of results yielded by value-added assessments. To this end three methodological issues were discussed. First, it was argued that simple average change or improvement on some measure of student development often provides a misleading estimate of the long-term effects of college. Alternative cross-sectional designs combined with statistical controls were suggested as ways to arrive at a more accurate estimate of the net effect of college.

Second, it was suggested that value-added analyses become increasingly useful in terms of policy as they increase their focus on the specific aspects of the collegiate experience which may affect student development. In addition to estimating the net effects of an institution, do they increase our understanding of how that institution functions to influence student development? To this end, causal modeling was suggested as an approach to value-added assessment which: 1) provided a theoretical template for understanding the process by which the effects of college or other educational experiences occur, and 2) permitted one to obtain a more complete estimate of the impact of various college experience measures by estimating both direct and indirect effects on student outcomes.

Third, it was suggested that value-added assessments that consider only the general effects of college disregard the very real possibility that not all students may benefit equally from the same educational experiences. Despite their usefulness in terms of a parsimonious estimate of educational effects, average differences or correlations may conceal as much as they reveal (Feldman & Newcomb, 1969). The more revealing question in terms of both policy and understanding of the complex dynamics of institutional impact is: what kinds of students change in what ways when exposed to what kinds of educational experiences?

Finally, it should be pointed out that the accountability movement and its attendant concern for assessing the student development outcomes of postsecondary education is likely to be with us for some time. Ewell (1986) has suggested that no less than 16 states have, or are currently in the process of developing assessment plans for estimating the outcomes of public postsecondary education. This suggests that a good deal of postsecondary education policy, and perhaps even funding, will be based

on the data we provide. It behooves us as educators and social scientists to maximize the credibility of our assessments of student development by collecting data under the most rigorous conditions possible and analyzing them in ways which increase our understanding the full range of specific institutional impacts.

References

Alwin, D., & R. Hauser. "The Decomposition of Effects in Path Analysis." *American Sociological Review*, 40, (1975), 37-47.

Anderson, J. & F. Evans. "Causal Models in Educational Research: Recursive Models." *American Educational Research Journal*, 11, (1974), 29-39.

Astin, A. "Undergraduate Achievement and Institutional Excellence." *Science*, 161, (1968), 661-668.

Astin, A. *Four Critical Years*. San Francisco: Jossey-Bass, 1977.

Astin, A. *Minorities in American Higher Education*. San Francisco: Jossey-Bass, 1982.

Astin, A. "Student Involvement: A Developmental Theory for Higher Education." *Journal of College Student Personnel*, 25, (1984), 307-308.

Astin, A. *Achieving Institutional Excellence*. San Francisco: Jossey-Bass, 1985a.

Astin, A. "The Value-Added Debate . . . Continued." AAHE *Bulletin*, 38, (1985b), 11-12.

Astin, A., & R. Panos. *The Educational and Vocational Development of College Students*. Washington, DC: American Council on Education, 1969.

Bean, J. "Interaction Effects Based on Class Level in an Explanatory Model of College Student Dropout Syndrome." *American Educational Research Journal*, 22, (1985), 35-64.

Beeken, L. *The General Education Component of the Curriculum through Transcript Analysis at Three Virginia Community Colleges*. Unpublished doctoral dissertation, Virginia Tech, 1982.

Blackburn, R., E. Armstrong, C. Conrad, J. Didham, & T. McKune. *Changing Practices in Undergraduate Education*. Berkeley, CA: Carnegie Council for Policy Studies in Higher Education, 1976.

Bowen, H. *Investment in Learning*. San Francisco: Jossey-Bass, 1977.

Buenz, R., & I. Merill. "Effects of Effort on Retention and Enjoyment." *Journal of Educational Psychology*, 59, (1968), 4-158.

Campbell, D., & J. Stanley. "Experimental and Quasi-Experimental Designs for Research." In N. Gage (Ed.). *Handbook of Research on Teaching*. Chicago: Rand McNally, 1963.

Cohen, J., & P. Cohen. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum, 1975.

Cronbach, L., & R. Snow. *Aptitudes and Instructional Methods: A Handbook for Research and Interactions*. New York: Irvington, 1977.

Domino, G. "Differential Prediction of Academic Achievement in Conforming and Independent Settings." *Journal of Educational Psychology*, 59, (1968), 256-260.

Dumont, R., & R. Troelstrup. "Measures and Predictors of Educational Growth with Four Years of College." *Research in Higher Education*, 14, (1981), 31-47.

Ewell, P. "The Value-Added Debate . . . Continued." *AAHE Bulletin*, 38, (1985), 12-13.

Ewell, P. *Personal Communication*. 9 May 1986.

Feldman, K., & T. Newcomb. *The Impact of College on Students*. San Francisco: Jossey-Bass, 1969.

Fincher, C. "What Is Value-Added Education?" *Research in Higher Education*, 22, (1985), 395-398.

Fincher, C. "The Emerging Role of Assessment and Evaluation in Postsecondary Education." Paper presented at the Third Annual Research Conference on Higher Education in Georgia. Athens, GA: 15-16 January 1986.

Hartle, T. "The Growing Interest in Measuring the Educational Achievement of College Students." In C. Adelman (Ed.). *Assessment in American Higher Education*. Washington, DC: Office of Educational Research and Improvement. U.S. Department of Education, 1986, pp. 1-11.

Heise, D. *Causal Analysis*. New York: Wiley, 1975.

Holland, J. "Explorations of a Theory of Vocational Choice and Achievement II: A Four Year Predictive Study." *Psychological Reports*, 12, (1963), 547-594.

Hyman, H., C. Wright, & J. Reed. *The Enduring Effects of Education*. Chicago: University of Chicago Press, 1975.

Lenning, O., L. Munday, & J. Maxey. "Student Educational Growth During the First Two Years of College." *College and University*, 44, (1969), 145-153.

Lord, F. "A Paradox in the Interpretation of Group Comparisons." *Psychological Bulletin*, 68, (1967), 304-305.

Maruyama, G., & H. Walberg, "Causal Modeling," In H. Mitzel (Ed.). *Encyclopedia of Educational Research*. 5th ed. New York: Free Press, 1982.

McCombs, B., & G. Smith. *The Effects of Differential Coursework on Student Learning in College*. Proposal submitted to the Office of Educational Research and Improvement, U.S. Department of Education. Denver: University of Denver, 1986.

McMillan, J. "Beyond Value Added: Improvement Alone Is Not Enough." Unpublished manuscript, Virginia Commonwealth University, 1986.

Mentkowski, M., & M. Strait. "A Longitudinal Study of Student Change in Cognitive Development and Generic Abilities in an Outcome-Centered Liberal Arts Curriculum." Paper presented at the annual meeting of the American Educational Research Association, Montreal, April 1983.

Mortimer, K., et al. *Involvement in Learning: Realizing the Potential of American Higher Education*. Report of the Study Group on the Conditions of Excellence in American Higher Education. Washington, DC: National Institute of Education, 1984.

Pace, C. R. *Evaluating Liberal Education: A Report Prepared for the Lilly Endowment*. Los Angeles, CA: Graduate School of Education, University of California at Los Angeles, 1974.

Pace, C.R. *Measuring Outcomes of College: Fifty Years of Findings and Recommendations for the Future*. San Francisco: Jossey-Bass, 1979.

Pace, C.R. *Measuring the Quality of College Student Experiences*. Higher Education Research Institute, Graduate School of Education, University of California at Los Angeles, 1984.

Pace, C.R. "Perspectives and Problems in Student Outcomes Research." In P. Ewell (Ed.) *Assessing Educational Outcomes*, San Francisco: Jossey-Bass, 1985.

Pallas, A., & K. Alexander. "Sex Differences in Quantitative SAT Performance: New Evidence on the Differential Coursework Hypothesis." *American Educational Research Journal*, 20, (1983), 165-182.

Pascarella, E. "Interactive Effects of Prior Mathematics Preparation and Level of Instructional Support in College Calculus." *American Educational Research Journal*, 15, (1978), 275-285.

Pascarella, E. "Student-Faculty Informal Contact and College Outcomes." *Review of Educational Research*, 50, (1980), 545-595.

Pascarella, E. "College Environmental Influences on Learning and Cognitive Development." In J. Smart (Ed.). *Higher Education: Handbook of Theory and Research*. New York: Agathon Press and AERA Division J, 1985a.

Pascarella, E. "Students' Affective Development within the College Environment." *Journal of Higher Education*, 56, (1985b), 640-663.

Pascarella, E. "The Influence of Living On-Campus Versus Commuting to College on Intellectual and Interpersonal Self-Confidence." *Journal of College Student Personnel*, 26, (1985c), 292-299.

Pascarella, E., J. Smart, C. Ethington, & M. Nettles. "The Influence of College on Self-Concept: A Consideration of Race and Gender Differences." *American Educational Research Journal*, (1980), forthcoming.

Pascarella, E., J. Smart, & J. Stoecker. "College Racial Composition and the Early Educational, Occupational and Economic Attainments of the Black Men and Women." Unpublished manuscript, University of Illinois, 1986.

Pascarella, E., & P. Terenzini. "Interaction Effects in Spady's and Tinto's Conceptual Models of College Dropout." *Sociology of Education*, 52, (1979), 197-210.

Pascarella, E., P. Terenzini, & L. Wolfe. "Orientation to College and Freshman Year Persistence/Withdrawal Decisions." *Journal of Higher Education*, 57, (1986), 155-175.

Perry, W. *Forms of Intellectual and Ethical Development in the College Years*. New York: Holt, Rinehart and Winston, 1970.

Pfeifer, C. "Relationship Between Scholastic Aptitude, Perception of University Climate and College Success for Black and White Students." *Journal of Applied Psychology*, 61, (1976), 341-347.

Plant, W. *Personality Changes Associated with a College Education*. U. S. Department of Health, Education, and Welfare Cooperative Research Branch Project 348 (SAE 7666). San Jose, CA: San Jose State College, 1962.

Prather, J., J. Williams, & J. Wadley. *The Relationship of Major Field of Study with Undergraduate Course Grades: A Multivariate Analysis Controlling for Academic and Personal Characteristics and Longitudinal Trends*. Atlanta, GA: Georgia State University, Office of Institutional Research, Report OIP-77-3, 1976.

Ratcliff, J. *The Effects of Differential Coursework on Student Learning in College*. Proposal submitted to the Office of Educational Research and Improvement, U.S. Department of Education, Ames, Ia.: Iowa State University, 1980.

Ross, S., & E. Rakow. "Learner Control Versus Program as Adaptive Strategies for Selection of Instructional Support on Math Rules." *Journal of Educational Psychology*, 73, (1981), 5-753.

Solmon, L., & P. Taubman. *Does College Matter?* New York: Academic Press, 1973.

Spaeth, J., & A. Greeley. *Recent Alumni and Higher Education.* New York: McGraw-Hill, 1970.

Stinard, T., & W. Dolphin. "Which Students Benefit from Self-Paced Mastery Instruction and Why." *Journal of Educational Psychology, 73,* (1981), 754-763.

Strange, C. *Intellectual Development, Motive for Education and Learning Styles During the College Years: A Comparison of Adult and Traditional Age College Students.* Unpublished doctoral dissertation, University of Iowa, 1978.

Telford, C., & W. Plant. *The Psychological Impact of the Public Two-Year College on Certain Non-Intellectual Functions.* Department of Health, Education, and Welfare Cooperative Research Branch Project SAE 8640. San Jose, CA: San Jose State College, 1963.

Terenzini, P. & E. Pascarella. "Student-Faculty Relationships and Freshman Year Educational Outcomes: A Further Investigation." *Journal of College Student Personnel, 21,* (1980), 411-418.

Trent, J., & L. Medsker. *Beyond High School.* San Francisco: Jossey-Bass, 1968.

Volkwein, J., M. King, & P. Terenzini. "Student-Faculty Relationships and Intellectual Growth among Transfer Students." *Journal of Higher Education, 57,* (1986), 413-430.

Warren, J. "The Blind Alley of Value Added." AAHE *Bulletin,* 37, (1984), 10-13.

Whitla, D. *Value Added: Measuring the Impact of Under-Graduate Education.* Cambridge, MA: Office of Instructional Research and Evaluation, Harvard University, 1978.

Wittrock, M. *Handbook of Research on Teaching.* 3rd ed. New York: Macmillan, 1986.

Wolfle, L. "Strategies of Path Analysis." *American Educational Research Journal, 17,* (1980), 183-209.

Wolfle, L. "Effects of Higher Education on Ability for Blacks and Whites. *Research in Higher Education, 19,* (1983), 3-10.

Wolfle, L. "Applications of Causal Models in Higher Education." In J. Smart (Ed.), *Higher Education: Handbook of Theory and Research.* New York: Agathon Press and AERA Division J, 1985, pp. 381-413.

An Assessment of Assessment

RUSSELL EDGERTON
President, American Association for Higher Education

I am flattered to be on this distinguished program—also awed by the audience and the assignment. I'm neither a testing/assessment specialist nor a foot soldier in the assessment movement. What I know comes from looking out my window, first at FIPSE, and now at AAHE, at the general flow of events in higher education.

So for courage, I turned to the inspiring words of Yogi Berra, who once said: "You can observe a lot by just watching."

For about 18 months my colleagues and I at AAHE have been watching in fascination as assessment surfaced into a major public issue. What we see is lots of confusion—confusion even about what assessment is.

Given this state of affairs, I thought I might first try to describe what the "it" is and what seems to be taking place around the country in the name of assessment. Then I'll try to draw back and comment on where we might go from here.

A Play of Four Acts

I've come to think of assessment as a play that has unfolded in four acts. In our current discussions, we tend to forget that the play has been running quite a while, so I'd like to balance this by concentrating especially on the earlier acts.

Act I

Act I is titled, "An Idea Is Born." Scene I takes place, not on campus, not even in America, but in Great Britain. The time is World War II.

The British were faced with the problem that the corps of officers to serve in the military had to be very rapidly expanded. Some method had to be invented for selecting who from the rank and file would make good officers.

To do this, selection boards were set up and candidates were put through a series of exercises. These included standardized intelligence tests and five other paper-and-pencil psychological tests. Each candidate was also interviewed by a three-person team consisting of a psychiatrist, a deputy from the selection board, and another officer.

But this wasn't all. The centerpiece experience was a series of standardized performance situations. The candidates had to do things like (1) interview someone playing the role of the candidate's subordinate, (2) give a "morale" speech, (3) take command of a group and lead it to accomplish a task, and (4) perform in a leaderless group situation, such as improvising an escape. At a final conference, all the people who had observed the candidate would get together to discuss the candidate's suitability for officer status.

In the summer of 1943, an American official from the newly created Office of Strategic Services observed all this and brought it back to us. OSS was faced with a task similar to that of the British: Who should be sent behind enemy lines?

OSS took the British methods and added a new ingredient. A team of Harvard psychologists led by Henry Murray was enticed to work on the problem too; their belief was that, by carefully observing how people behaved in various situations, one could infer underlying patterns of personality. By systematically varying the nature of situations individuals were put in, it was possible to construct a "picture of the whole person."

By war's end, Murray's theories and all the wartime experiments had come together into a new tradition called "assessment." As Georgine Loacker from Alverno College has reminded us, the word has Latin roots: *ad* plus *sedere* means, literally, "to sit down beside."

This ends Scene I. Techniques like the leaderless discussion group and the in-basket exercise gradually seeped into industry and government. But by and large these new methods were pooh-poohed by psychologists and testing professionals. The tradition might have died were it not for Douglas Bray, a researcher at AT&T.

Act I, Scene II, opens at AT&T in the mid-fifties. Bray was hired to establish a new program of basic research on managerial careers. To obtain baseline data for his studies, he established what he called an assessment center and put samples of AT&T employees through 3 1/2-day assessments at these centers. And—you guessed it—he demonstrated that assessment methods were smashingly successful at identifying who AT&T's successful managers would be.

Word filtered out through industrial psychologists, and soon major

corporations were starting assessment centers. Bray's scientific methods reinforced the intuition of corporate leaders that the abilities being measured were the ones that really counted in managerial success. In the 1970's the assessment center concept took off.

Thus, an idea was born that would soon travel to higher education.

I've told this story at length because it helps underscore a fundamental point about assessment. Oliver Wendell Holmes once said that "a word is but the skin of a living thought." The thought that lives inside the word "assessment" is a different thought from the one that lives inside words like "test" and "testing."

First of all, assessment focuses beyond what an individual *knows* to the abilities brought into play as the individual actually *performs* a task.

When we apply for a driver's license, we take a "test" on our knowledge of the rules of the road. Then we go out and give a sample of our ability to drive—or at least parallel park. The first is a test of knowledge, the second an assessment of competent performance.

(Granted, the sample is not what it might be. *Very few* people each year are killed parallel parking.)

Assessment also has a second, crucial characteristic. At its core, assessment is a methodology of *multiple* judgments. Candidates are put in *various* situations. For the "purists," at least one of these situations should be a simulation—like the in-basket test, where individuals work for several hours on a variety of things in their in-box, and then explain why they did what they did. Expert judgments of various assessors are *pooled* before a final conclusion is reached.

Thus, a single paper-and-pencil test, a single panel interview, or even a single performance excercise—when there is no pool of observations to judge—is not, technically speaking, an assessment.

This ends Act I. In the next three acts, three different groups of actors come onto the stage. They all borrow from this tradition and carry on work in the name of assessment. But this is about all they have in common. Under the label assessment, they are really pursuing quite different ends.

Act II

In Act II of the play, assessment is brought onto the college campus. The time is the 1970s, though there are flashbacks to earlier periods. The actors are principally individuals who work directly with students—faculty and counselors on the cutting edge of various movements in educa-

tional reform.

They belong to different departments, usually aren't aware of each other's work, and don't think of themselves as pioneers of a new movement. What they share is a common interest in pushing higher learning beyond a focus on teaching and evaluating what students *know* to teaching and evaluating what they can *do* — that is, to a focus on abilities required for effective performance.

One scene takes place in admissions and counseling offices on campuses that began dealing with large numbers of returning adult students. Some way is needed to help these adults translate their checkered educational backgrounds and informal learning into a currency that can be awarded academic credit. So these actors start asking their students to develop portfolios that can document their knowledge and competence. The Council for the Assessment of Experiential Learning is created to set standards for this process.

In another scene, faculty in occupational and professional [illegible]ems struggle to define the competencies required for effective perfo[illegible]e in their fields. They are bothered by the lack of fit between what's taught in the classroom and what counts on the job. They turn to the assessment tradition to bring the two worlds closer together.

The field of teacher education offers the latest example of a professional field turning to assessment. The "Holmes group" and the Carnegie Forum have recently issued major reports, calling for the transformation of teaching into a full-fledged profession. Carnegie has called for the creation of a new national board that will develop standards and procedures for entering the profession. Look closely at the initial vision for this process, developed by Lee Shulman and Gary Sykes of Stanford, and you'll see — not testing — but assessment. Candidates for teaching certificates will not simply take written tests but undergo full-fledged, 2-1/2-day "assessments": in-basket exercises, video-taped samples of actual teaching, "the woiks."

In the final scene in Act II, actors in liberal arts colleges and general education programs also turn to assessment. In 1973, Alverno College in Milwaukee decides to take abilities out from the shadows of the assumed curriculum and make them into an explicit "second curriculum." To help assess whether their students are indeed acquiring these abilities, they establish — with the help of AT&T — an assessment center. Others don't go as far as Alverno. But the idea that abilities like critical thinking, communicating, problem-solving, and making value decisions are central, and should be explicitly taught, slowly flows into the bloodstream. The

AAC report, *Integrity in the* … this idea the seal of approval.

In the course of these … comes to be *used* in a new way. Assessment d… ethod of selection -- a way to judge who sh… how assessment was used in WW II. It was …

But many of the e… …essment in the 1970s didn't want merely … …ance. They wanted to improve it. They aske… …such as learning how to learn are as importa… …now are they taught? The answer they came to … …east they aren't taught in the conventional sense … Students learn *about* things by being told; they lea… …y doing them.

John Dewey said it … …e that students cannot be taught many of the things they … …—but they can be *coached*. The role of the teacher, then, … …re .e conditions and design tasks through which student… …rough performance—and then comment on this performance as it goes along. That is, to be *assessors*.

To the players in Act II, then, assessment is a crucial dimension of effective teaching. Rather than view tests and exams as an afterthought—things that get administered at the end of a course because a record needs to be kept—they view them as a way to give students feedback on their performance.

A final scene in this act takes place inside the medical profession. Years ago, McMasters University in Canada converted to a curriculum that focused learning around real problems of medical practice. Rather than stuff students full of knowledge for two years and then place them in clinical practice, they start their students off, from day one, working in collaborative teams, solving problems, acquiring abilities of lifelong learning.

Last year, Harvard began an experimental program, called New Pathways, based on this approach, and this has boosted its national visibility. Assessment, in all this, is the key. FIPSE's round of grants for assessment, just announced, includes grants to the University of New Mexico's school of Medicine to establish a medical self-assessment center and to five New England universities to implement skills assessments for fourth-year medical students.

Act III

After an intermission, Act III opens on a very different note. The scene shifts to state capitols around the country. The time is the late 1970s and early 1980s. The central actors are legislators and governors, and campus leaders who deal with them.

All of you are familiar with the particular chain of events that aroused interest in the quality of our schools and colleges. Attention focuses first on the need for a larger pool of talented students in math and science. Then, with the release of the report, *A Nation at Risk*, this turns into a generalized concern about the quality of high schools. After several years of school reform, concern spills over into how things are going at the college and university level.

Even before the quality of undergraduate education became a popular issue, some states began worrying about students being admitted to college—and advancing—who seem to lack basic academic skills. Understandably, the states turn to the tools they used to enforce new standards on the schools: mandated tests. New Jersey begins the parade in 1977-78 with a statewide program of testing entering college students for basic skills. Georgia follows with a basic skills test for "rising juniors." Florida mandates both. Tennessee steps forward with a required entry-level test and a new twist, financial incentives for institutions to assess their two-year and four-year outcomes.

Much of this is done in the name of assessment. But needless to say, *this* version of assessment, motivated primarily by a concern for maintaining minimum standards of entry and academic progress, is a far cry from the assessment tradition we saw earlier. Minimum competency testing often relies on a single test as a basis for judgment. The test results are used for selection and comparison—as hurdles for students, and, further, as a basis for drawing comparisons among institutions.

As this act unfolds, these early state initiatives draw lots of fire. The presidents of Educational Testing Service and the College Board, to their credit, both speak out against overrelying on standardized tests to enforce minimum standards. They argue that:

- the focus on minimum competencies will skew priorities away from the more advanced knowledge and abilities that are also the concern of colleges;
- one shouldn't rely on a single test for advancement to matriculant or junior status, any more than one should rely just on a SAT score to determine admission to college;

- one shouldn't take performance scores on standardized exams and use these to infer things about the quality of programs students passed through. The exams were simply not designed for this purpose.

Evidence, especially from Florida, about the effect of minimum competency testing on minorities gives these arguments special poignancy and force.

Two years ago, educators feared that versions of the "Florida model" might spread to other states. But because of arguments like those just cited, and the effects of a minimum competency testing on minorities, the pressure seems now to have eased. Impressive governors such as Thomas Kean of New Jersey and John Ashcroft of Missouri have chaired thoughtful studies about how states should act to bring about quality improvement. The state-level conversation about assessment is now framed by their reports.

Peter Ewell of NCHEMS, who scouts the state policy terrain better than anyone I know, reports that the states now seem to be sorting themselves out into three basic camps.

First, there are states that lean toward the Florida end of the scale and have, or are seriously considering, statewide mandated tests of various kinds. A few of these states (New Jersey, Texas, and Maryland come to mind) have backed off the idea of rising-junior exams, and are refocusing their attention on entry testing and longer-term developmental work on outcomes.

Second, there are states that are setting statewide objectives for assessment, but permitting institutions to develop their own ways of finding out and reporting on how these objectives are being met. The state might say, we want you to report on student achievement in general education—you figure out what particular instruments to use.

A third category of states is requiring that institutions undertake concrete, regular investigations of educational outcomes in relation to their own instructional goals. In effect, they are saying: Show us that you have a good assessment process in place.

In all cases, however, the states are saying: We're tired of looking at college quality in terms of faculty/student ratios, library books, buildings and equipment, and other resources. We want you to measure college quality in terms of actual student learning.

One last point: Some states are also putting money on one or two institutions to be guinea pigs. In some cases, the initiative comes from the state; in others, campuses have stepped forward on their own and the

state has recognized their leadership. So now, Missouri has its Northeast Missouri State; Virginia has James Madison University; New Jersey has Kean College; Colorado has Colorado State; and New York has Plattsburg and Empire State—all engaged in programs of assessment.

Act IV

In Act IV, the stage shifts back to the campus. But the actors are not the original innovators who first led the assessment movement. This time, they are vice presidents, deans, directors of offices of institutional research, department chairs, and others concerned about evaluating academic programs and the effects of the overall campus experience.

They feel vulnerable to charges that they really don't have much evidence about how their students are really performing. Some believe that it's important to take preventive action: "If more assessment is to be done, better us than someone else." Others have more authentic educational interests, and concede the point: We *should* know more about how our students are doing.

Those that have thought hard about these matters also concede that our present evaluation activities fall short in two crucial respects.

Most campuses have offices of institutional research. In the early days, these offices evaluated curricula, assessed student achievement, predicted student success, compared different teaching methods, and studied student attitudes and satisfaction. But in the 1960s, with the arrival of computers and new pressures for managerial information, the focus of most of these offices shifted. They now produce reports on student flow, costs per credit hour, faculty salaries, surveys of teaching loads, and a dozen other topics deemed important for planning and budgeting—but unrelated to student learning.

Consequently, we know a lot about how students are administered, but very little about whether they are learning anything. The promise of the early efforts at institutional research has gone unfulfilled.

Second, the researchers who *are* asking important questions about student learning are doing so on a separate and independent track from the decision makers on and off campus who might use the results. As Derek Bok points out in his new book, *Higher Learning*, research studies on the impacts of the undergraduate experience have rarely come about at the request of faculty members or deans. The questions researchers have asked are not necessarily the questions that deans and faculty members

want answered—or governors.

There are, as always, notable exceptions. This past year, the comprehensive assessment programs of Alverno College, the University of Tennessee-Knoxville, and Northeast Missouri State—all described in Peter Ewell's timely book, *The Self-Regarding Institution*—have become highly visible national models. But three out of 3,000 isn't a very big number.

Now, a new generation of efforts is starting up. Even the prestigious universities and hot colleges are feeling the need to be able to talk about their quality from a position of strength—that is, with evidence. At Harvard, Derek Bok has asked Richard Light, an expert on evaluation from the Kennedy School of Government, to organize a new project. Each month several dozen key faculty and administrators are getting together, joined by guests from other campuses, to ask questions about issues like:

- Writing. What is the concrete evidence that we are really teaching this ability well? What do we know about what works?
- Instruction. What methods are in place for giving faculty "fast feedback" on their effectiveness?
- Programs like New Pathways. How can these approaches be evaluated?

Act IV is still going on. In my version of the last scene, all the actors come out onto the stage and move around, passing each other. At first they pass each other without recognition, then notice each other but aren't sure what to do. They turn and look to the center, waiting for leadership.

A Pop Quiz

So much for the play. Now for a pop quiz.
Assessment is:

A. A specific tradition of evaluation based on theories and methods distinct from those that underlie tests and testing.

B. Part of a new movement toward more active teaching and learning, that focuses on performance, multiple judgments about this performance, and feedback to the student.

C. The label for a new state insistence that campuses maintain minimum standards and demonstrate what students are actually learning.

D. The label for a renewed campus interest in collecting information about student performance that is useful for evaluating the effectiveness of programs and the overall campus experience.

E. All of the above.

You got it. It's all of the above—which is why the play is at once confusing, depressing, and exciting.

A Better Script

Well, what about it? Should the play go on? Actually, it *will* go on, no matter what the reviews are in *The Chronicle of Higher Education*. What we can influence is the quality of the performance.

Daniel Yankelovich has noted that public issues tend to go through cycles. First there is a consciousness-raising phase, a time when we become aware that a problem exists. Then there is a phase that he calls the "working through process" when the issues become clarified, the implications understood. Finally there is the phase when we settle into a common judgment that informs a course of action.

Campuses and states are now strung out along various points of the working-through process. A handful of campuses and a half-dozen states, early leaders in the movement, have made up their minds about what should be done, but most have not. This means that there is time to do things right.

The question is, how are we going to use this time? At the moment, the campus engagement with assessment is a mile wide and an inch deep. Are we going to regard assessment as an issue to finesse—to dispose of as quickly as possible? Or are we going to take it to the level it deserves—and deal with the deeper questions about our educational purposes?

If the latter, each of the actors in the drama has some challenging questions to work through.

The States

Take the states. The first question the states must face up to is: Do we want our campuses simply to document student performance or really improve it?

In the negative scenario, the states will simply lay on new layers of tests and reporting requirements, and misuse the data collected for

simplistic institutional comparisons. (Given the penchant for statistical measures of quality control and America's love of score-keeping, we kid ourselves if we think that data, once collected, won't be used—*and misused*—for comparisons.)

In the positive scenario, the states would say: "You've been too smug (which is true); we don't want you to go on assuming that your students are learning; we want you to collect and confront evidence on the point; and we want you to do so in a way that the information comes actually to be used to improve the education you provide."

A second question for the states is: What *level* of performance do we have in mind?

To date, the campus dialogue with the states has turned on minimum standards. It's like the dialogue that started in the late 1970s between the automobile industry and the larger public. When the automobile industry turned out cars that had defective brakes, doors that rattled, and so on, the public got fed up and demanded improvement.

Our own equivalent to cars with defective brakes are students who accept public student-aid funds and then don't show up for class; athletes who are carried along without meeting minimum expectations for academic performance; and graduates who cannot write clear sentences. The public has a right to feel outraged about these things. We should too.

But all this defines a quality product as one that is *absent of obvious defects*. We should all hope that the public expects more of our students and our colleges, and we expect more of ourselves, than this!

The difficulty, of course, is that higher standards are harder to agree on. It's easy to agree on minimums. But as we aspire to higher levels, we run into the problem that standards can be set only in the context of some definition of what objectives are most important to us. If we want to purchase a car that is not simply free of obvious defects, but performs at a higher level, we have to answer another set of questions: Do we care most about gas mileage, in which case we'd buy a small, light car; or safety, in which case we'd buy a heavier car? Lacking such judgments, we can't really evaluate what we mean by higher performance.

So it is with students. What learning during the college years do we care most about? Are we satisfied if our students acquire enough specialized knowledge to find a professional, technical, or managerial job? Or do we want them to have the capability to improve their jobs, and learn new jobs over their careers? And how about attitudes such as tolerance of people different from them, or feeling responsible for others?

A third question is: Do we want to improve performance, whatever the

cost, or do we see the quality movement in a context of *overall* social needs?

The shortest and easiest way to improve quality, in one view, is to raise admissions standards and let in only those students who are academically gifted and easy to teach.

But we can't maintain our democratic way of life on into the next century by simply focusing our educational energies on the top 25 percent—or even 50 percent—of high school graduates. Everything I read about work in the new economy and citizenship in the information age suggests that we must *broaden* the base of talent as well:

- We are going to earn our national living, not because old-fashioned entrepreneurs establish new organizations along Taylorist principles, but because *whole organizations* are dynamic—that is, staffed with managers and workers who are enterprising and capable of change. As Rosabeth Kanter puts it, the new entrepreneurship consists of "thousands of little battles of people with initiative."
- Twenty years ago, I used to teach my political science students that political information flowed in a two-stage process. Party leaders and other elites would pass the cues about what to do along to the rank and file. Now, courtesy of Dan Rather, every citizen learns what Reagan said to Gorbachev at the same time as Bob Dole and Tip O'Neil, and we don't wait for *them* to make up our minds.

In brief, shortcut strategies to quality improvement don't meet our real needs. We have to *expand* access *and* improve quality at the same time.

The Campus

Whether or not the states are willing to be firm but flexible, in turn, depends on whether campus leaders take assessment as a serious issue and then move it in positive directions.

Peter Ewell reports that many campuses now see assessment as something to be undertaken for the sake of appearances. Committees are being set up simply as a response to external pressures. Their work is not connected to curriculum planning or review, budget planning or review, or any other decisions that count.

A negative scenario is not hard to foresee. Assessment, so managed, will become another bureaucratic add-on. In addition to taking course examinations, students will have to take a new battery of exams adminis-

tered by officials in some remote office. The faculty will regard all this as yet another plot by administrators to hassle their lives.

Alternatively, campus leaders could view the emergence of the assessment issue as a timely moment to reengage serious and basic questions. What are our students really learning? What are they like when they leave? How can we improve our contribution to their development?

The good news, from those campuses that have taken this path, is that it is good for one's health. At Alverno, Northeast Missouri State, and Tennessee-Knoxville, the interest in assessment started as an administrative initiative. But the faculty, initially suspicious, were brought into the process and soon assumed ownership of it. Assessment turned up evidence that has prompted real program improvements. The process has enhanced their visibility and prestige. Enrollments are up. And assessment has justified increased funding.

The cautious news is that each of these models came about through a long, evolutionary process. The stimulus and the support came over a decade ago. The Tennessee Commission on Higher Education, the stimulus for the UT-Knoxville's work, started working on a performance-funding project in the mid-1970s. Alverno introduced its curriculum in 1973. Northeast Missouri has been at this about the same period of time. The quick fix just isn't in the cards.

Testing and Accrediting Agencies

Much will also depend on the attitudes taken and roles played by the testing agencies and other organizations that provide essential products and set standards for quality.

Testing and assessment require expertise. Colleges need help to do it right. But the kind of help they need, in the first instance, is not a test but a *service*. They need someone to talk to who can say, "If you want to measure this learning outcome, or that program, then here are the tools that are available." They need primers and guides for how to get started, how to clarify and put values on the outcomes they want to assess, and how to frame issues in the most productive way.

In a context, then, of understanding and authoritative advice, colleges will need new instruments for measuring student performance. But they don't need full-course, take-it-or-leave-it meals so much as menus and access to cafeterias of products, in various price ranges, from which they can select what's most useful—products they can build into their own,

home-grown assessment programs.

Some of these products might be standardized tests, but not necessarily the ones now on the shelf — or at least not in the package they now come in. The difficulty with present instruments is that they are designed for purposes of selection. Given this purpose, the tests are constructed so as to spread out student differences for the purpose of relative rankings. The ideal test item is one which 50 percent of the students answer correctly, and 50 percent fail.

The answers on this sort of test can't tell us what students know or can do in relation to some criterion of actual performance, but simply on how well they perform relative to the others taking the test. Accordingly, these answers are not very useful for the purpose of giving feedback to students, or for evaluating the quality of the academic program they experienced. Indeed, in the case of a number of standardized tests, students and institutions can't even obtain the information about what they answered right and wrong. The scores they receive are scores derived from the performance of all the test takers.

I'm just a country lawyer, but it seems to me that testing agencies are now in a new era. It's an era that calls upon agencies such as Educational Testing Service to rethink its mission; to consider becoming, in name and self-concept, Educational Testing *and Assessment* Service; to move beyond the development of selection tests to the development of measures of real attainment; and to think not only about how to develop products but how to provide services to campuses and states struggling with issues beyond their comprehension.

Accrediting agencies face an even more challenging set of questions. Everywhere I turn — whether to presidents, state officials, or the research community — I find a growing sentiment for a major overhaul. But this is a topic for another day.

A Common Agenda

I want to turn to one final theme. Besides facing up to the deeper issues, we need to find a common agenda. The campuses can't do it right acting simply on their own. Yet outside agencies can't be expected to invest resources in new developmental programs unless campuses and states clarify what the priorities are.

All of higher education doesn't have to be interested in assessment in order to entice some big players into the game. But those who are

interested do have to work together to define some collective needs. So, we return to the quandry of what—beyond minimum standards—is important to assess.

This is no time to take the lid fully off this box, but I do want to make some parting comments about the direction we should all point our compasses.

First, which way is north? Much of the conversation about undergraduate reform in the past few years has looked inward and back, to the golden era of the 1960s. If only we could return to the days of yesteryear when students arrived with high SAT scores; when they were socially concerned and enrolling in liberal arts courses; when the curriculum was coherent; when faculty salaries were high, and so on. The baselines for studies by Astin, Bowen, and many others are rooted in that era. We see ourselves as having fallen from that high estate.

Whatever one's view of those times, the reality is that they are gone, and the conditions that made it possible will never return. North is not inward and backward, but outward and forward. We need to derive our expectations for what students should learn from the requirements of the future.

Second, it's a future in which the requirements for specialization, for knowledge in depth, will intensify. We should take it for granted that our students will need to pursue more and more specialized subjects.

The challenge here, for teaching, testing, and assessment, is to move along the path outlined in T. S. Eliot's "Choruses from 'The Rock'": "Where is the knowledge lost in information? Where is the wisdom lost in knowledge?" The only way I know to help students deal with spreading complexity and accelerating change is to help them grasp more and more of the underlying pattern of things.

Students who rushed into computer science in the last 10 years are going to wish, in another 10, that they had gotten a better grounding in basic mathematics. In a world of change, only what's fundamental endures. And there's nothing more fundamental than a high-powered theory.

I worry, therefore, that higher standards and more testing might simply add to the already heavy pressures for that easiest of matters to test, subject-matter coverage. As Charles Muscatine has remarked, for many students, "taking a course is like trying to get a sip of water from a fire hose." Students are deluged with information. I've never met a faculty member who didn't believe that most of what he or she knew was indispensable knowledge for everyone.

We need, then, to discipline ourselves, to test and assess our students for their grasp of only the most fundamental and key concepts. Less is better than more. Knowledge is better than information. Understanding is better than knowledge.

Third, most all of us would agree that our students should acquire more than simply specialized knowledge. We're very unclear about what the stuff of this remaining agenda is.

I think we are on the verge of a new construct, a new way to conceive and think about the breadth requirements of undergraduate education. The old construct was a construct of general and liberal knowledge. It was based on the Renaissance ideal of a person in touch with all the basic fields of knowledge. It made great sense at the turn of the century. It makes little sense today.

The new construct is not built around a commitment that a given body of knowledge is essential. Rather, it is built around the view that certain abilities, attitudes, and perhaps even certain social skills are essential to effective performance in a wide variety of roles. These can be acquired in the course of both specialized and general studies.

Given this view, I very much hope that the assessment movement can steer around the sterile debate over general and liberal knowledge. Some campuses may want to follow Secretary Bennett's old-time religion and return to a classic curriculum. Others may follow a different path. Our view should be, as Senator Howard Baker used to say, "I ain't got no dog in that fight."

Whatever the context for subject-matter knowledge, we all should be able to agree on the higher-order abilities that are required in a society of spreading complexity and accelerating change. This is the first and foremost agenda for collective action. It's what the thoughtful elements of the general public most expect college graduates to have. It's what the original promise of the assessment tradition is all about.

Suppose, for instance, that several consortiums of institutions got together and committed themselves to the objective of turning out graduates who could *communicate* what they know effectively. This would build upon the acknowledged importance of writing and the writing-across-the-curriculum movement, but take it to a new and more complex ability level. It would include the intellectual acts of synthesizing information, as well as the sensitivities and techniques of writing, listening, and speaking.

Faculties within the participating institutions, in many disciplines, would pledge to teach and assess their students for this ability. Students

taking coursework would be asked to demonstrate what they know, not only on paper-and-pencil tests, and not only in writing, but also by communicating their ideas in various social situations. ETS or other agencies could prepare guides for how to design these situations, how to train assessors, and what to look for.

One final point. Let's remember that while abilities like communicating effectively are a *double* agenda for the near future, our long-term common agenda includes other characteristics as well.

As I look to the future, everything I see suggests that four clusters of characteristics are becoming ever more important:

- The first is captured by words like initiative, persistence, drive, and risk-taking;
- The second, by words like flexibility, resourcefulness, and openness to change;
- The third, by words like character, integrity, and responsibility toward others;
- The fourth, by words like leadership and effectiveness in working with others.

As I look to the environments from which our students are coming, everything I see suggests that families and other institutions that used to foster these characteristics have weakened, while other institutions that erode these characteristics, like the mass media, have grown in strength and influence.

So, the big question lying ahead is, what role will colleges and universities play in this? Are we going to drift along passively with the dominant culture and simply offer students an assembly line of courses? Or are we going to muster the internal convictions and public support to play a counterweight role in American life, to develop educational processes that are empowering, educational environments that affirm important values, and thus turn out students who will create the kind of society we want to live in?

When we pass out grades and tests, we make decisions about these deeper values. We now distribute grades and scores as if students were in a contest with each other, a contest in which many compete but few are victors. Yet we increasingly serve students who don't need to lose more races; they need to be showered with praise for things really mastered. And the world they will enter requires not only competition but cooperation.

Why do we hold our students accountable for individual performance

when the world demands group performance? Why do we give grades in the first place? What should we make of the fact that some of the colleges that have thought most deeply about these issues, and have the most intensive and comprehensive programs of assessment—institutions such as the McMasters University Medical School and Alverno College—don't give grades at all?

Let us hope, in short, that the assessment movement will bring us to reflect, more deeply than we otherwise would have done, on what we are trying to accomplish.

Thank you.