DOCUMENT RESUME

ED 434 570 HE 032 348

AUTHOR	Miller, Margaret A.; Ehrmann, Stephen C.; Rohan, Sue; Eaton, Judith S.; Mac Gregor, Jean; Wergin, Jon F.; Alberts, Bruce
TITLE	Architecture for Change: Information as Foundation. 1998 AAHE Assessment Conference (Cincinnati, OH, June 13-17, 1998).
INSTITUTION	American Association for Higher Education, Washington, DC.
ISBN	ISBN-1-56377-044-X
PUB DATE	1998-06-00
NOTE	95p.; Foreword by Barbara L. Cambridge.
AVAILABLE FROM	American Association for Higher Education, One Dupont Circle, Suite 360, Washington, DC 20036-1110. Tel: 202-293-6440.
PUB TYPE	Collected Works - Proceedings (021)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Educational Assessment; Educational Attitudes; Educational Change; *Educational Quality; Elementary Secondary Education; Evaluation; *Higher Education; Institutional Evaluation; Outcomes of Education; Program Evaluation; Quality Control; Science Education; Student Evaluation; Teaching Methods
IDENTIFIERS	Malcolm Baldrige National Quality Award; Quality Assurance

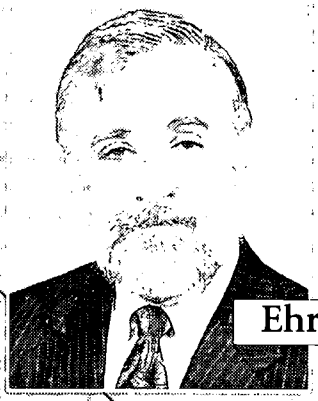
ABSTRACT

The seven speeches in these proceedings were presented at the 1998 American Association for Higher Education conference on assessment. "Blueprints" sets out the architecture of the conference with four "strands," each focusing on one level of information. The first is assessment of pedagogies in the classroom, on campus, and beyond; the second is assessment of programs and units; the third is assessment within and across institutions; and the fourth is moving from information to action. Four keynote addresses, each addressing one strand, follow: "What Outcomes Assessment Misses" argues that assessment is a way to clarify what we want to achieve, how we can get there, and how we can know when we have arrived; "Assessment of Powerful Pedagogies: Classroom, Campus, and Beyond" examines assessment of alternative teaching methods, such as collaborative, active, and problem-centered learning; "The Malcolm Baldrige Approach and Assessment" discusses the Malcolm Baldrige National Quality Award approach to self-assessment; and "Assessment of Programs and Units" reviews the challenges of effective program assessment. A plenary speech, "Accreditation and Quality Assurance: Ambivalence and Confusion," discusses quality assurance through accreditation; another, "Reinvigorating Science Education in the U.S.: The Importance of Appropriate Assessments," discusses how appropriate assessment can help improve the poor state of math and science education in the United States. (MDM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 434 570

Miller



Ehrmann

June 13-17, 1998
Cincinnati, OH

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

B. Pollack

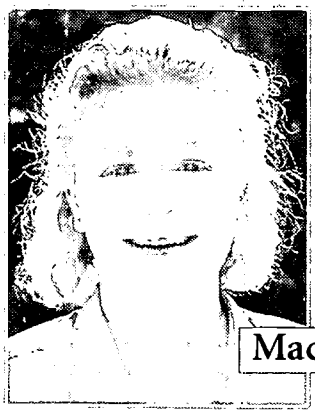
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1998 AAHE ASSESSMENT CONFERENCE

Architecture for Change: Information as Foundation

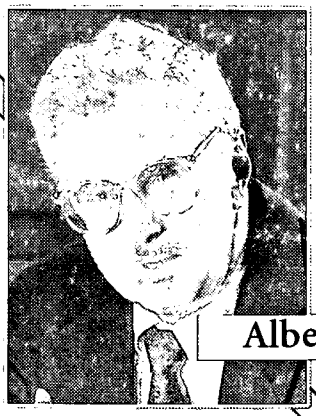
Rohan

Eaton



MacGregor

Wergin



Alberts

HE 032 948

June 13-17, 1998
Cincinnati, OH

1998 AAHE ASSESSMENT CONFERENCE

Architecture for Change: Information as Foundation

Ehrmann
Miller
Rohan
Eaton
MacGregor
Wergin
Alberts

Architecture for Change: Information as Foundation
Presentations from the 1998 AAHE Assessment Conference

© 1999 American Association for Higher Education. All rights reserved. Printed in the United States of America.

Opinions expressed are the contributors' and do not necessarily represent those of the Association or its members.

Ordering Information
Additional copies of this publication can be purchased by contacting:

American Association for Higher Education
One Dupont Circle, Suite 360
Washington, DC 20036-1110
(ph) 202/293-6440, (fax) 202/293-0073, (email) pubs@aahe.org, (website) www.aahe.org

ISBN 1-56377-044-X

ERIC

Contents

Architecture for Change: Information as Foundation

Foreword v
by Barbara L. Cambridge

Blueprint 1
by Margaret A. Miller

Strand Keynotes

What Outcomes Assessment Misses 7
by Stephen C. Ehrmann

Assessment of Powerful Pedagogies:
Classroom, Campus, and Beyond 27
by Jean MacGregor

The Malcolm Baldrige Approach and Assessment 43
by Sue Rohan

Assessment of Programs and Units 59
by Jon F. Wergin

Plenary Speeches

Accreditation and Quality Assurance:
Ambivalence and Confusion 67
by Judith S. Eaton

Reinvigorating Science Education in the U.S.:
The Importance of the Appropriate Assessments 73
by Bruce Alberts

Foreword

Architecture for Change: Information as Foundation

by Barbara L. Cambridge

The book *Metaphors We Live By*[1] has stimulated my thinking since I first encountered it years ago. It points out that metaphors are ways that we structure our experiences to make sense of them. When planners of the 1998 AAHE Assessment Conference chose the theme of "Architecture for Change," for example, we wanted to provide in the conference program everything from a firm foundation to specialized features so that all conferees could build on their own unique experience and knowledge.

The analogies and metaphors that central speakers at the 1998 conference used to construct portions of their talks point to important messages of the conference and of their presentations. Jean MacGregor spoke for AAHE goals when she described her role in assessment work: "a bit Perle Mesta (convening conversations), a bit Johnny Appleseed (traveling around picking up and planting seeds of good ideas), and a bit Saul Alinsky (organizing on behalf of institutional change to support innovation and reform efforts)." The presentations that you will read in this collection bespeak persons who in quite varied settings, from professional association to governmental agency to college campus, play Mesta, Appleseed, or Alinsky at different times. Their roles mean that they take up different metaphors to explain their work and the work of all of us who use assessment to learn and to improve.

Margaret A. Miller, president of AAHE, set the scene at the conference with an introduction to its four thematic strands. She noted that assessment, which pays attention to results, is now threaded through most of institutional life. Quoting a poet, Miller stated about assessment: "Everything is stitched with its color." Indeed, great progress has been made over the past decade in incorporating assessment practices into the fabric of institutional life, through classroom assessment, program review, accreditation, and institutional representations to multiple publics.

Yet, knitting those practices into whole cloth continues to be a challenge. In fact, four speakers spoke of current concerns. In introducing the strand on program reviews, Jon Wergin warned that these reviews are too often "one-shot affairs," not well integrated into the life of the institution. Encouraging fidelity to our reasoned programs and practices, he

[1] George Lakoff and Mark Johnson (Chicago: University of Chicago Press, 1980).

calls on faculty members and administrators to "put our strong academic values of systematic inquiry and questioning of assumptions to use" in a continuous review process. In a plenary presentation, Judith Eaton identified a common concern about accreditation. She describes it as baggy: "big, elastic, inefficient." These characteristics of a voluntary system of accreditation are less negative however, Eaton says, when we consider the alternative of a government-operated system. Eaton speaks about ways that the Council for Higher Education Accreditation and faculty members at all colleges and universities can contribute to a vital accrediting process that serves multiple stakeholders in higher education. Sue Rohan acknowledges this range of key stakeholders. Sometimes the needs of students, faculty members, taxpayers, employers, and governing boards compete: How to use the standards of the Baldrige National Quality Award to work toward doing well what each of these groups needs is the gist of Rohan's presentation. In another plenary talk, Bruce Alberts recounts the way in which the setting of academic standards in science, a "hot potato," landed in the hands of the National Academy of Sciences, which he heads. Blistering standardized tests, Alberts calls for authentic assessment to move math and science students toward deeper learning. Each speaker at the conference identified through figurative and descriptive language the issues that face us all.

Although these and other challenges sometimes seem overwhelming, Steve Ehrmann contends in his strand introduction that assessment presents a way to clarify what we want to achieve, how we can get there, and how we can know whether we have arrived. Evoking the illuminating focus of a flashlight in the night, Ehrmann calls his current work the Flashlight Project. Assessment in his metaphor can help in "spotting an elephant in the dark." Indeed, if this book sheds light on effective assessment practices for you, it will have fulfilled its purpose. ●

.....

Director of the AAHE Assessment Forum at the time of the 1998 Assessment Conference, Barbara Cambridge now directs the AAHE Teaching Initiatives, including the Carnegie Teaching Academy Campus Program.

Blueprint

by Margaret A. Miller

When I introduced the first assessment conference in Virginia in 1986, I laughed when I said, "Welcome to what the planners of this conference want me to refer to as 'the first' annual Virginia Assessment Conference." It seemed to me at the time that assessment was a relatively straightforward affair — a program that decided what learning it wanted to accomplish, analyzed its effectiveness, and used that information to improve itself. How hard could that be? Ten years later, at the tenth annual Virginia Assessment Conference, I was no longer laughing.

Assessment turned out to be technically much more difficult to do than we had anticipated, at least in reliable, valid, and subtle enough ways. More important, assessment required a kind of self-reflexivity that constituted a remarkably profound cultural shift for the academy. To some degree, and in some places, the cultural shift has happened. On some campuses, attention to results now threads through institutional life — where "everything," as the poet says, "is stitched with its color!"

But at this point in the history of higher education in America, all campuses must systematically produce and examine evidence of their effectiveness and use that information for improvement and decision making. The theme of the 1999 AAHE National Conference on Higher Education, which will be held in Washington, DC, March 20-23, will be "Organizing for Learning: Core Values, Competitive Contexts." As we see it, the chief challenge now for higher education is to prepare students for life, work, and citizenship in a complex and interconnected world, and to do that job in such a way as to preserve our fundamental values in an increasingly competitive higher education environment. The foundation of that work will be information: information to improve what we have done traditionally; information to monitor the effects of experimentation, change, and variation in pedagogies, programs, and institutions; and information to support choice and decision making.

Four Strands, Four Levels

Hence the "architecture" of this conference. Each of its four strands focuses on one level at which information is vital.

The first level is the classroom, although

Assessment turned out to be technically much more difficult to do than we had anticipated, at least in reliable, valid, and subtle enough ways.

the "classroom" is increasingly becoming wherever the student happens to be, on or off campus. At this level, faculty need answers to the deceptively simple question of how well the teaching strategies they use generate learning. That question is at the heart of two new AAHE projects. One, in cooperation with the Carnegie Foundation for the Advancement of Teaching and funded by the Pew Charitable Trusts, centers on the scholarship of teaching. The other, a project on science education reform, funded by the National Science Foundation, concentrates on institutional reform to support teaching and learning.

The second level at which assessment needs to occur is the program. The pragmatic need for information at this level is generally for purposes of formal review, often motivated by external forces such as specialized accreditation or state-mandated program review. More important, this kind of assessment focuses on the culmination of a student's entire educational experience.

Assessment at the institutional level probably motivated many of you to come here today — at least those of you who are dealing with regional accreditation. But information at this level is also crucial to good campus decision making about what to continue doing, what to stop doing, and where to put resources.

If you're at a public institution, you may also need institution-level information to satisfy outside entities. Coordinating boards and legislatures need this information, not just to hold institutions accountable for the expenditure of public funds but also to make decisions about what to support. With clear answers to questions about institutions, students might choose among the bewildering variety of postsecondary options on better grounds than price tag, reputation, or the look of the place on the day of a campus visit. And employers with hiring decisions to make could use good information about what graduates know and can do.

Let me say a few things more about each of these strands.

Strand One:
"Assessment of Powerful Pedagogies: Classroom, Campus, and Beyond"

We need information on the classroom level and at the level of teaching strategy. Several things strike me as I look at the pedagogical innovations of the past decade.

First, we educators ask more of ourselves, as we should, given the demands students face when they graduate. We want students to learn deeply, and we want to engage their hearts as well as their heads. We want them to be changed by their education.

Second, our expanding knowledge of how people learn has greatly increased our repertory of strategies. For example, we know how important cocurricular activities are to student understanding. We know the value of service-learning, which places students in volunteer situations that provide a real-life context for what they're learning in class. Service-learning addresses both our desire to produce graduates who are good citizens and our understanding that students who participate in service-learning situations are more apt to learn. As Andy Clark has put it, perception in human beings is not a contemplative affair; it is a "recipe for action" (*Daedalus,* Spring 1998, p. 267). Some innovative pedagogical strategies have come out of assessment itself. Being

clear with students about what teachers expect them to learn turns out be a powerful learning tool; so too are the student self-assessments that are built into many of the best assessment programs.

Third, we have new technological tools to use to good and even transformative effect, as The TLT Group, AAHE's teaching, learning, and technology affiliate, keeps reminding us. All of these changes require that we become sophisticated scholars of teaching and learning; that we carefully and precisely trace the effects of those strategies on students' understanding. In the first strand of this conference, you will see some of the best of that scholarship displayed.

Strand Two:
"Assessment of Programs and Units: Program Review and Specialized Accreditation"

The second strand reaches directly into the felt self-interest of many faculty and staff, who have a deep sense of ownership of, and sense of community within, their departments and programs. A good deal of the evaluation with real consequences for the future of the unit focuses on this level; program review and specialized accreditation particularly. The challenges in this area are several.

The first challenge is to coordinate the information gathering that is required of the various processes such as accreditation and program review. The second is to ensure that the information produced is valid, reliable, and subtle enough to drive important decision making. This is one of the biggest technical challenges of assessment. The third challenge is to adapt the call for information to the unit's own purposes: to answer questions faculty actually have about students, to find out what they need to know to improve the program. And the final challenge is to actually use the information — to connect the program to the larger institution, to build it, to improve it, and to stop doing what doesn't work.

Strand Three:
"Assessment Within and Across Institutions: Institutional Effectiveness and Regional Accreditation"

I said that many of you are probably here today because you are facing a regional accreditation visit. Over the past thirteen years, that is probably the single most compelling reason for people to come to an AAHE Assessment Conference. The regionals began asking for evidence of institutional effectiveness in about 1985. But in recent years, state policymakers have been asking for it as well, often in directive and reductive ways such as through performance measures. Assessment moves here beyond the institution's boundaries and becomes as much a matter of accountability as of improvement. Although it's hard not to feel defensive when this happens, the key is to use these pressures for the institution's own ends.

What are some ways in which a trans-institutional perspective can be helpful? First, accreditation is a time to make sure that your institution has integrity — that its values and its results line up, that changes made on campus are consistent with those values, and that the institution is not just a collection of programs and activities but a coherent whole.

Second, the recent move on the part of many states to institute performance measures suggests that they want indicators of institutional effectiveness. If the ones set for you seem wrong, I'd encourage you to ask yourselves by what measures you would, as a campus, be willing to be held accountable?

AAHE has another new project, in partnership with Indiana University Purdue University Indianapolis, in which six public comprehensive urban institutions will develop institutional portfolios. Those portfolios will contain the evidence of results by which these institutions measure their own effectiveness. It should be a model of how institutions can take charge of their own self-definition in the face of others' attempts to define them.

Finally, there is no way to know whether an institution is successful at something without a context — as we say in my family, it's all a matter of "compared to what?" Benchmarking your results and processes against those of a like institution can tell you where you are doing better than expected and where you have something to learn.

Strand Four:
"Information to Action: Asking Good Questions, Generating Useful Answers, and Communicating Well"

In the last strand, we come to the cornerstone of this conference and AAHE's notions of assessment. Here, we explore what it means to ask good assessment questions. One of the stories I used to tell my students was about the Nobel physicist Isidore Rabi, whose mother used to ask him when he came home from school not "What did you learn today?" but "Did you ask a good question today, Izzy?" As we all found out when we became professionals, the capacity to ask a good question is what separates the expert from the novice. So the first challenge we address in strand four is what kinds of questions to ask. For instance, Steve Ehrmann, in the Flashlight Project, which assesses the effects of technologically delivered instruction, suggests that when we're comparing online against live courses, we might want to ask not "How does the learning in the two compare?" but instead "What different kinds of learning do they generate?" By the way, in producing information for a particular audience, such as students, it's probably a good idea to ask them what kinds of questions they actually have about colleges and universities.

The second challenge is to use the information. If we're not prepared to change teaching strategies and programs on the basis of what we learn, or better support them, or even terminate them, I'd suggest that assessment is a sterile activity doomed to languish in a campus corner.

Finally, in this strand too we move beyond the institution's borders to consider its place in a larger context. How can we communicate honestly, precisely, and comprehensibly to higher education's many supporters what we are doing well and what we are not? What kinds of information do they need to make decisions about where to go to college, how to distribute resources, and how to hold us accountable?

Finally

My husband, Alan Howard, who runs a Web-based American Studies program at the University of Virginia, has said that he watches commercials the way some people look at the faces of their sleeping children — alert to the flickers that might give a clue about the dream going on beneath the surface. He recently pointed to one commercial that has intrigued him.

In the commercial's first frame, a CEO addresses a group of suits, exhorting them to "think outside the box." Cut to the basement, where we see boxes moving on conveyor belts. The boxes are moving quickly and efficiently, and the message seems to be that in the box-moving business, what's

needed is a faster and more efficient way to do that work. But no one here is actually questioning the boxes themselves — what's in them, whether they're the right size and shape, or more radically, whether this company is actually in the box-moving business at all. If it's in the goods-moving business, maybe the boxes aren't even needed.

Alan likes this commercial as a metaphor for the instructional uses of technology, but it works for assessment too. We can use assessment to do better the things we have always done. There is considerable virtue in that — courses and programs and institutions constitute the structures in which most of us, and our students, now live, and we need to be sure that they serve their purposes well. But assessment now has a more intriguing role.

We live in a world in which traditional, mass higher education is faced, perhaps for the first time, with serious competition. As Ted Marchese wrote in a recent *AAHE Bulletin* article on the new education providers (May 1998), we're now living in a world where "everybody goes after the other guy's lunch." That's the bad news.

The good news is that competition drives innovation, and the need for innovation takes us back to first principles, territory we should revisit every once in a while. The discipline of innovation, according to Peter Drucker, begins with knowing what business you're in. Innovation also makes assessment essential in a way it isn't when it's business as usual. Drucker describes the next three imperatives of innovation as assessing your results, abandoning what doesn't work, and assessing again.

If we're not prepared to change teaching strategies and programs on the basis of what we learn, or better support them, or even terminate them, I'd suggest that assessment is a sterile activity doomed to languish in a campus corner.

The alternative higher education providers not only stimulate us to innovate. By the examples the best of them set — from program development that begins with learning goals, to the habit of continuously assessing and improving their programs — they also challenge our ways of working. They even raise the fundamental question of the business we're in.

I'd suggest that although our assessments might be organized at the classroom, program, and institutional levels, we're not in the course, program, or even campus business. We're in the learning business — that's the goods we need to deliver, maybe in classrooms and programs and on campuses, and maybe not. We need to assess whether we're generating that learning, change or abandon strategies to do so that don't work, and reassess continuously.

AAHE's role, to quote from our mission statement, is to help "institutions develop their capacities to make the organizational, pedagogical, and other changes needed to achieve their evolving missions." With that purpose in mind, I welcome you to the latest of our baker's dozen, the thirteenth AAHE Assessment Conference. ●

Margaret Miller is president of the American Association for Higher Education. For fifteen years, she was an English professor then campus administrator at the University of

Massachusetts at Dartmouth. In 1986, she moved to the State Council of Higher Education for Virginia, where she served between 1987 and 1997 as chief academic officer. At the Council, Miller worked with faculty and academic administrators and was responsible for the approval, review, and assessment of academic programs throughout Virginia.

At AAHE, Peg Miller continues to work toward the organization's goals of bringing together thoughtful constituents to address the major challenges currently facing higher education in turbulent times: how we can organize for and assess learning, support and evaluate teaching, extend education beyond the classroom into the community, deal with changing faculty roles, use the new technologies responsibly, ensure quality, communicate our results to the public, and level the speed bump between K-12 and collegiate education.

What Outcomes Assessment Misses

by Stephen C. Ehrmann

A Summary, in Advance

A speaker benefits from having an easy straw man to knock over. Here's mine. If you're going to evaluate a program, common wisdom says to:

- Assess the educational outcomes of that program (only).
- Look at how well the average student achieves those goals.
- Develop your tests and inquiry so that, ideally, you will be able to report achievement rather than being forced to look at and talk about failure.

I'm going to try and knock over all three of those contentions, to argue that each one of them is radically incomplete as a way of looking at our programs of instruction. The problems they share have particular relevance to the uses of technology, but the problems are also important to the study of almost any educational program.

First, I'll argue that evaluation is more than just a matter of outcomes assessment. Although the fourth principle of good practice in assessment reminds us to look at students' experiences, not just at what they learn, the commonplace view seems to be that assessment can begin and end with the question "Did they learn it?" I'll try to point out some of the benefits of attending to means, not just ends.

Then, in a clever little pun, I'm going to shift from means to the mean — that is, the average. I'll talk about the crucial information that is missed when we look only at common goals and average scores, especially in programs that use technology to expand creative work and work on open-ended problems.

In the third and final segment of this talk, I'll argue that good news can be hidden in bad news, that patterns of persistent failure can yield fresh insight into a program's most dearly held values, and that this kind of evaluative data can provide a

foundation for a fresh approach to faculty development.

I. The "Means" Matter

Ends matter, but so do means. If we don't study *how* a result was achieved (as opposed to the way we planned to achieve it), data about *whether* the result was achieved is not very useful.

The simplest form of this argument is really easy to make. Imagine that we evaluate a program only by comparing its outcomes with something else, for example with the program's performance last year or with the outcomes from a competitor program. The data show that the program could be performing 10% better, let's say. Without some insight into what people actually did in the program (as opposed to what they said they would do behind those closed classroom doors or while off doing homework), how can we decide what to do next to improve those outcomes? Since learning is most directly the result of what students do, studying what students actually did in a course, as opposed to what we hope or fear that they did, yields useful information.

How can typical faculty members and administrators look at *process* — at the means — in ways that complement outcomes and that can guide changes in policy and practice?

Asserting Some Definitions

That's a big question, but before answering it, I'm overdue to assert some definitions. I say "assert" because none of the following terms has widely agreed upon definitions, so it's my responsibility to say what I mean by each of them.

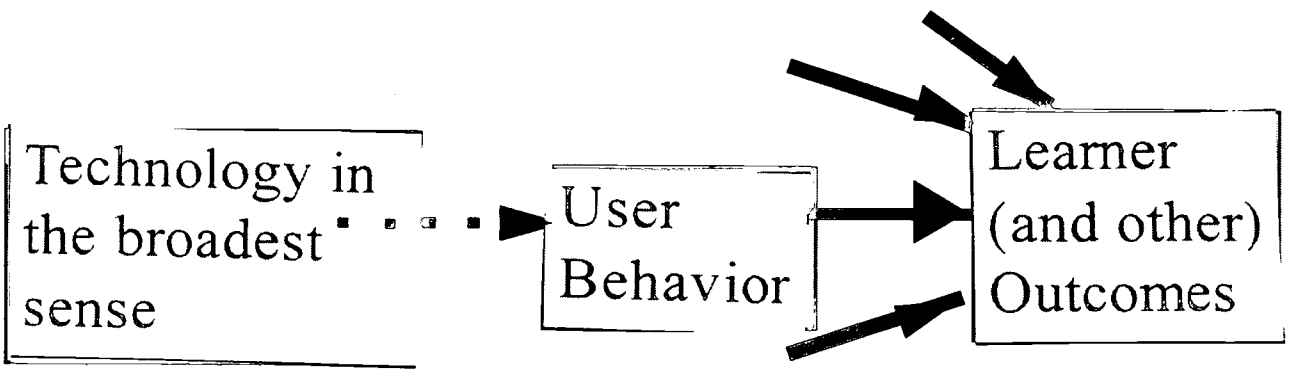
Figure 1 *(opposite)* sets up a relationship among technology, user behavior, and learner and other outcomes. On the left-hand side of Figure 1 is a box representing the technology of the program, which includes not just computers but chalkboards, the campus, and the way that faculty are organized — that is, the hardware, software, and social technology of the situation. The middle box represents what people chose to do with the technology. The right-hand box is the outcomes of what they did. For example, our technology, right here and right now, includes this lecture hall and me; that's the left box. The "users" of the technology are you; you're choosing to pay some degree of attention and some of you are taking notes; that's the middle box. If someone were to test you later on what you remember or what you've done as a result of this talk, those are the outcomes, the right-hand box.

In addition to *technology, user behavior,* and *outcomes,* I need to clarify some other ambiguous terms. When I say *assessment,* I mean measuring the outcomes included in the right-hand box. When I say *evaluation,* I'm talking about inquiry into how well the three boxes are functioning together — Are users doing what was expected with the technology (and, if not, why not) and, if so, are the desired outcomes occurring (why or why not)? So assessment produces information that is crucial for evaluation.

When I say *learning,* I'll be talking about the middle box, the user behavior. And when I talk about *learning outcomes,* I'll be talking about the right-hand box. So, usually when I use a phrase like "teaching and learning," I'll mean what teachers and learners are doing right now (not students' learning outcomes).

Notice some other relationships among the boxes in this figure. First, a dotted line from technology to user behavior reminds

Figure 1


Definitions
Technology in the broadest sense
User Behavior
Learner (and other) Outcomes
Assessment: outcomes
Evaluation: studying the program
Learning -Outcomes

us that the user has choices about what to do with the technology and that technology is not the only determinant of user behavior. What users do with technology is often not what the teacher or designer assumed and hoped that they would do. That's one proof that the technology is indeed empowering!

Second, lots of arrows go into the outcomes box, not just the line that goes from technology to user behavior. Whatever the user does with the technology is only one influence on the outcomes. For example, how much did the users already know before the intervention started? Because so many other factors can affect outcomes, it's risky to reason purely from outcomes data about how to change technology or behavior.

"You Idiot ..."

"You idiot," people have occasionally said to me (using politer terms, I'll admit). "It's simple to figure out the importance of technology using only outcomes data. You just do a controlled experiment." They claim that it's possible to learn all we need to know about the outcome by studying only the right-hand box, if we are very careful about how we make the comparison. A controlled experiment into the role of technology occurs when we set up two versions of a process that are identical except for the technology.

But how often can faculty members do an experiment that's so carefully designed that the design can rule out all extraneous factors and enable valid inferences about the technology's distinctive role? For example, how can typical experiments control what the students do? Although controlled experiments may be possible in big research studies, we're talking about evaluation of what is being done here and now, not about research that focuses on averages in multiple sites. Tip O'Neill once said that all politics is local. I hope we can agree that "all education is local." What happens on the average (research) tells us only a very little about what is happening to us (evaluation). Most of the factors leading into the right-hand box are very context-specific, very much about what's happening right here, right now, this year, with these people. If we can't "control" for variations in student motivation and talent, in precisely what the faculty member does, and in what's going on in the rest of students' studies, outcomes cannot tell for sure whether the technology itself worked.

As if that weren't enough, there's a second difficulty in relying only on outcomes data to make sense of technology. We would like to compare outcomes of two methods we have used, Method 1 with Method 2, in order to decide which is better, or whether we're making progress. But can we directly compare outcomes? What if the faculty member took advantage of the technology to change the goals of the course in Method 2? After all, one common reason to use technology is to help change *what* is being learned.

For example, consider a course in statistics (or graphic arts, or any of the other courses whose content is intimately tied to the use of some technology). Method 1, let's say, is a statistics course of study taught thirty years ago with paper and pencil methods. Because students could use only paper and pencil (and maybe a simple adding machine) to do homework or tests, the course of study could teach only certain statistical techniques and certain ways of thinking about data. The assignments, quizzes, and exams fit that vision of the course.

Method 2 is a contemporary course in

which students use graphing calculators, powerful computers with graphical displays, and huge statistical databases on the Internet. Because the field itself and the available tools have changed dramatically, faculty have made major changes in what they want students to learn. The course of study is now organized around different kinds of statistical techniques. Students also learn different attitudes and approaches to dealing with data, approaches that are more iterative, more visual. And, of course, the tests of achievement are dramatically different from those of thirty years ago.

So, if the tests of achievement are differ ent for the experimental group, Method 1, and the comparison group, Method 2, we cannot compare average test scores — outcomes — to decide how valuable the computers are. Let's stick with our statistics example. Let's assume that the average score of 78% on the final exam is the same in the experimental group and the comparison group. Other outcomes measures such as job placement rates and student satisfaction are also unchanged. Because we know that computers are currently important for learning marketable skills in statistics, we have to conclude that a simple comparison of outcomes is producing inadequate, even misleading, results.

If Comparing Outcomes Is Inadequate, What Do We Do?

I suggest two solutions. We can do better with the assessment comparison than my example suggested above, so we'll begin there. Then I'll return to the basic problem, which even the following suggestion doesn't totally resolve.

For the statistics course, we can produce a more productive result by comparing tests as well as test scores. We can ask a panel of judges whose judgment we trust — employers, graduate school representatives, faculty members who teach the courses that have statistics as a prerequisite — to examine not just the scores but also the tests themselves. We can ask them to choose Method 1 with its tests and test results or Method 2 with its tests and test results, considering, of course, the cost of teaching each method. Judges can report which method they prefer and why. That process is one way out of this quandary about outcomes.

To improve a course of study, faculty members usually need information on *how* the technology was actually used to complement whatever outcomes data or inferences about outcomes that they gather.

But we still have a problem: Just knowing that respected judges preferred the computer-supported course doesn't tell us enough to enable further improvement of the course and advances in cost efficiency. Although we know the results of the course, we still know very little about how the results were achieved, even in a course we taught ourselves, because so much depends on what students did when we couldn't see them and on what they were thinking at the time.

To improve a course of study, faculty members usually need information on *how* the technology was actually used to complement whatever outcomes data or inferences about outcomes that they can gather.

Looking at the Mean

A second solution to our problem involves looking at the mean. After identifying educationally important practices (the middle box in Figure 1) that depend on the technology, we can select practices we suspect can make the difference between good outcomes and bad. For example, we

might consider the seven principles of good practice, Gamson and Chickering's answer to the question "What does research tell us are practices that usually lead to good learning outcomes?"[1] If we wanted to explore the value of technology, we might find that some of the seven principles (e.g., student-faculty interaction and active learning) were implemented more thoroughly in Method 2 and that the technology was being used by students in their active learning and their interaction with faculty.

Finally, *if* we were unable to measure directly whether outcomes were better than for a comparison (perhaps we're studying Physics 101 and Physics 103), it would still be interesting to know whether one group's use of technology was helping them implement the seven principles of good practice better than the other group was implementing them. These seven principles are so important because there's so much research showing that implementing these kinds of practices yields better learning outcomes.

For example, imagine you're in an institution that has spent a lot of money on email and Internet connectivity. Your institution wants to educate students who are better skilled at working in teams than graduates were a decade ago. Further, you may have data showing that graduates of a program are getting better at working in teams, but you'd still like to know whether the email had anything to do with that improvement. A necessary step is determining whether the email was used by students to work in teams. How often did they use it? Are different types of students, such as commuting students or students whose native language is not English, using email more than other types of students? Are some kinds of students benefitting more or less than the norm? When trying to work in teams, did students find the email a real help, or did they make their teams work despite barriers posed by the email media and the email system? Answers to those questions and others like them would help to show what, if anything, your email investment had to do with the improvement in outcomes.

Suppose that you found that email was *not* being used effectively to support improvements in the skills of graduating students. Then other questions might occur to you. How about the training for using email for this purpose? How reliable is the system? How often do students use their computers for other purposes (that might affect how often they log on)? How reliable is the email service?

By getting answers to these questions you begin to build up a story of the role that the technology is playing or failing to play in supporting the strategies in which you are interested.

To sketch technology's role in helping students learn, you can address at least five types of questions, four of which are *not* outcomes assessment. The first three correspond to the three boxes (the *triad)* in Figure 1.

1. Questions about the technology, *per se* (e.g., Could students get access to it? How reliable was it? How good was the general training? Are some students more familiar or skilled with the technology than others?).
2. Questions about the practice or behavior, *per se* (e.g., How often are students asked to work in groups? What training do they get in team skills? Are some

[1] The basic principles were first laid out in Chickering and Gamson (1987). They are repeated, and the use of technology in supporting them is explored, in Chickering and Ehrmann (1997), which is also available on the Web.

students already good at this coming into the program?).

3. Questions about the outcome, *per se* (e.g., What changes are there in team skills of graduating students? How often are they called on to use those skills after graduation? How well do they do in those settings?). This is where outcomes assessment fits.

Then we have two more sets of questions, about the arrows:

4. Questions about the technology's use for the practice (e.g., How satisfactory was email as a medium for team work? How often was email used for that purpose?).
5. Questions about the practice's fostering of the outcomes (e.g., Did commuting students who rate high in group skills also work extensively via email? Do graduates interviewed about their work in groups talk about group work they did in college that involved email?).

It turns out that many different disciplines and types of institutions are using technology in similar ways, for similar reasons, and with similar anxieties. That's what makes the Flashlight Project possible and useful.

This project, which I direct, has been developing and distributing survey and interview questions of these five types. Many Flashlight items focus on the seven principles of good practice, the ways that students and faculty use technology to implement those principles, and some of the most common problems that can block the functioning of such triads. Information about Flashlight is available on the website of The TLT Group at <www.tltgroup.org>. If you click on "FLASHLIGHT" in the table of contents, you'll find material, including a summary of the issues and technologies we currently cover ("The Flashlight Project: Spotting an Elephant in the Dark").

The site also includes links to Flashlight-based research reports, such as one by Gary Brown, at Washington State University. Brown's report provides an example of using Flashlight to study how an outcome was achieved in an experimental seminar program for at-risk students at WSU. Higher GPAs indicated that the students coming out of this program were probably benefitting, but had technology played a role?

Armed with Flashlight data about student learning practices, Brown and his colleagues developed a convincing story about how the freshman year gains were achieved: Technology was being used to implement principles of good practice. These findings were used as part of a successful argument to institutionalize the program.

II. What the Mean Misses

The second part of my straw man focuses on "what the mean misses." When I was at the Evergreen State College in the late 1970s, I served as director of educational research. As part of my job, I would periodically ask a faculty member how I could help in doing evaluations. I'd say, "You pick the question. I will provide all of the money and half the time needed to answer the question. You will need to do the other half of the work. So, if you want to find out something, let's work together on devising a really good question."

Faculty often replied, "OK, what's a good question look like?" I would answer, "Imagine your program as a black box. A mass of students is marching in one end of the box and some time later they come out the other side, changed. How do you want

them to be different after the program is over? Once you tell me that, we'll see if we can come up with a 'difference detector' that is very carefully geared to noticing whether this change has, in fact, happened, and we'll go on from there."

I quickly discovered that there were three kinds of faculty at Evergreen. One sort of faculty member enthusiastically and decisively answered how students should be different, and we went on from there. A second group answered my question "How do you want the students leaving to be different from the students entering?" rather more hesitantly. They had an answer, but they and I weren't too satisfied with it. Finally, a third group couldn't answer my question at all: They couldn't say how they wanted students to become different as a result of their program. So I concluded, being 27 at the time, that this was the difference between very good faculty, mediocre faculty, and faculty who really didn't know what they were doing.

I then moved to the Fund for the Improvement of Postsecondary Education (FIPSE), where one of my duties was to work with applicants and project directors on their evaluation plans. I would ask them the same question: "How do you want people to be changed as a result of being encountered with your project?" Amazingly, FIPSE project directors fell into the same three categories of great FIPSE project directors, mediocre directors, and directors who never should have been funded in the first place. Except that categorization was clearly ludicrous. Many of these projects were clearly superlative, despite the fact that my categories slotted them as directionless. But if they were so good, why couldn't they answer this seemingly simple question: "What do you want the average student to learn as a result of his or her encounter with your project?"

It took me some years to see the difficulty. My question had presumed a particular goal that was uniform for every student: some particular way in which students were all to be changed by the program. Figure 2 *(opposite)* helps illustrate my presumption. In Figure 2, each student is represented by an arrow. Students' knowledge before entering the program is represented by the base of the arrow — some know more than others at the start. The tips of their arrows represent their capabilities by the end of the program. We can see that they learned different amounts, but (we assume) they all learned the same *kind* of thing — the only thing we're concerned about — learning in line with the program's educational goal.

I now call this the "uniform impact" perspective on education, because the educator's goals are what count: These goals are the same for all students, and a good program impacts even students who initially don't want to learn. It's a very legitimate, logical way to look at education. But, as you know, it's not the only way to look at education.

Figure 3 *(below Figure 2)* offers a second perspective. It presumes that the educational program is an opportunity. Different people come in with different needs and different capabilities. Accidents and coincidences happen. Students are creative in different ways, leading to still more diversity of outcomes from the same course or experience. After the program, former students move into different life situations, further changing the shape of the program's successes and failures. In short, for many reasons, different people learn different things as a result of their encounter with a learning opportunity. These differences in

Figure 2

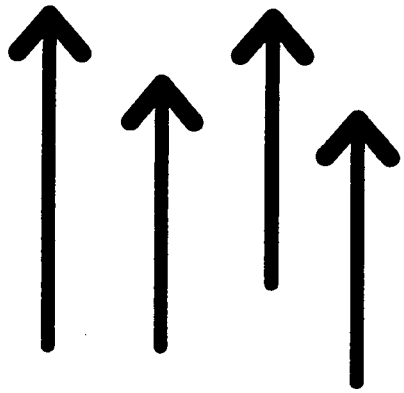
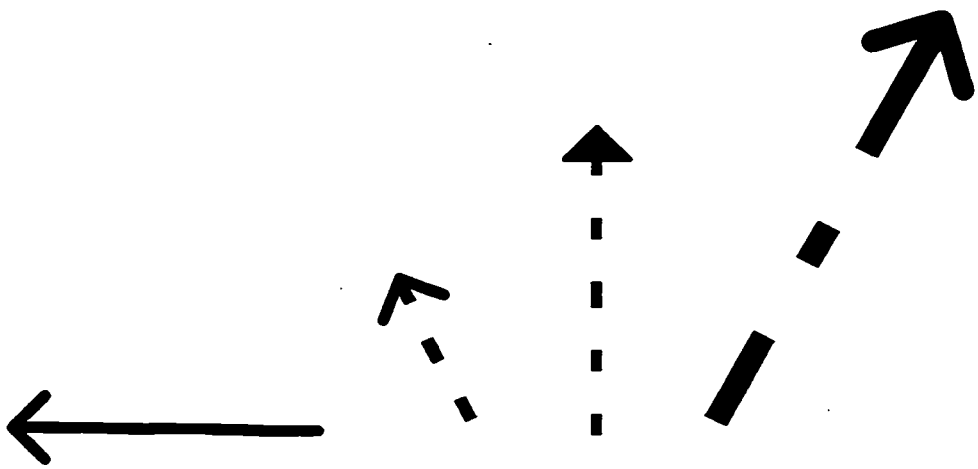


Figure 3



learning are qualitative, that is, different in kind; and quantitative, that is, different in degree.

Figure 3 might represent all four people in a very tiny English class. One masters grammar, one becomes a great poet, one falls in love with Jane Austen's novels, and one picks up skills that eventually lead to a job in advertising. Imposing a uniform impact perspective labels the course a failure. If its goal was to teach poetry, the average learner became only slightly more interested. If its goal was to teach grammar, ditto. Almost no one learned about Jane Austen. And so on. But if the goal was that learners took away something of life-changing importance related to English, the course was 100% successful.

These qualitative differences in learning can sometimes be quite big from one learner to the next, especially if the instruction is meant to be empowering, research-oriented, exploratory, individualized. And, of course, learner empowerment is often the intent of using computers and telecommunications.

I call this perspective "unique uses," because it begins with the assumption that learners are unique and that we are interested in how they've made use of the educational opportunity that is facing them. The key to assessing learning in unique uses terms is not whether students all learned some particular thing (uniform impact) but rather whether they learned something — anything — that was quite valuable (by some broad, multi-faceted standard or process we use for determining value.) In the English class of four students described above, the unique uses criterion used was whether the learning was of life-changing importance and whether it had something to do with English.

College effectiveness ought to be viewed mainly from the unique uses perspective, especially in the liberal (liberating) arts. What, *on the average,* is a college supposed to achieve for its liberal graduates? College-wide learning goals are difficult to agree upon if restricted to specifying what all learners must learn: the lowest common denominator of geography majors, literature majors, and physics majors. On the other hand, if the goal for graduates is (also) that something terrific happens to them as a direct result of their college education, *no matter what that outcome is,* we will notice very different things about their learning and their lives. We might notice that two members of one graduating class won Nobel prizes, for example, and credited their undergraduate educations in their acceptance speeches, even though we'd never put "winning a Nobel prize" on a list of uniform impact goals for undergraduates.

Each perspective — uniform impact and unique uses — picks up something different about what's going on in that single reality. This is not, in other words, a case of the good new perspective versus the bad old one. In almost any educational program these are two quite legitimate ways to assess learning and to evaluate program performance. Each focuses on elements that the other tends to ignore.

When designing any assessment or evaluation, the relative importance of those two perspectives is going to depend on the educational program itself and the client's needs. For a training program, a uniform impact perspective might catch virtually everything of interest to a policymaker: Did every doctor in the program master that particular open-heart surgical operation? On the other side, evaluating the educational performance of a university may warrant relatively modest attention to the uniform impact perspective. Most of the important outcomes differ in kind from one

department to the next and from one student to the next. Usually, however, both perspectives are required to do an assessment or evaluation that is fair and reasonably complete.

As teachers, we apply both perspectives all the time. We want students to master subject-verb agreement, so with subjective, expert judgment we design a test of that skill. The students' scores signal (we hope) whether they have a deep, lasting understanding of subject-verb agreement. That's uniform impact assessment. We may evaluate the course's performance each year in this area by the average scores of students on the test.

In the same course, we also assign the theme "What characteristics of a college course help us learn?" We give the resulting papers to an external grader who grades each essay — A, B, C, D, C, C, B, A. If we ask the grader, "Why did you give those two papers B's?" or "What did those three C papers have in common?" the answer might well be, "They have nothing whatever in common, those three C papers, except they were all C work." That's a unique uses assessment. The grader had different reasons for assigning each of the A's, each of the B's, and each of the other grades. We might then ask the grader, "How good is this year's version of the course?" And the grader (if he had graded essays for this course before) probably would have an opinion. That opinion might also include an expert judgment on how good the course was in stimulating a variety of types of good writing. That's a unique uses evaluation.

That's just what happened at Brown University in a study of the use of a precursor of the World Wide Web (see Beeman et al. 1988). As was customary for this English course, Professor George Landow used an external grader on the essays for his experimental section. The grader had years of experience grading final exam essays for this course, and when she was shown the essay questions in advance she told Landow what he might want to consider. "This will be a very difficult essay test," she warned. He said, "No, no, that's all right. I want to give this test." The external grader must have agreed in the end that students performed well on the test, because she gave many of the students A's. There was probably a great diversity of achievement among those students, different kinds of excellence, because of the web of resources and the manner in which Landow had taken advantage of that web in organizing the section's work. So, after assessing each student's excellence, the grader drew an evaluative conclusion: excellent course.

The next point of distinction between the uniform impact perspective and the unique uses point of view is their *contrasting definitions of excellence.*

Through uniform impact lenses, we see excellence in the ability to produce the desired goal. One approach is better than another if it's better at adding value in that particular direction and can do so consistently even in a somewhat different setting and with different staff. The term "teacher-proof" is one variation on this theme: The program produces results even if teachers aren't especially good. For example, a calculus program is wonderful because even when students come in hating calculus, they love it by the end of the program, and their scores on calculus achievement tests are really high. In uniform impact terms, this is a wonderful, wonderful program.

To determine whether a program is excellent in unique uses terms, on the other hand, evaluates the magnitude and variety of the best achievements of the students,

Each perspective — uniform impact and unique uses — picks up something different about what's going on in that single reality. This is not, in other words, a case of the good new perspective versus the bad old one.

after assessing the students' work one at a time. Judging a program *design* as excellent involves asking how many *different* ways it has been adapted to different settings and produced appropriate excellence.

Here's an example of the recognition of the importance of variety. In 1987, I was involved with one of the first large-scale uses of "chat rooms" in composition programs. The approach, developed originally by Professor Trent Batson, of Gallaudet University, was called the ENFI Project, Educational Networking for Interaction. Faculty in the project, to some extent, did their own thing, embroidering on the basic ENFI motif. But shouldn't they all be doing the same thing if the evaluation was to mean anything? Batson had, after all, been funded to test the practice of chat rooms in multiple settings.

For better or worse, nonetheless, faculty were using somewhat different technology, and somewhat different teaching methods, thereby exercising their academic freedom with a vengeance. The uniform impact puzzle then was, "Are they all doing 'ENFI'?" From a unique uses perspective, however, we could ask, "Has the concept of ENFI stimulated each faculty member to do something wonderful and effective for his or her students?" In fact, it would be a mark of the strength of the ENFI concept if *different* adaptations of the ENFI idea usually worked, even if in different ways (see Bruce, Peyton, and Batson 1993).

For me, by the way, Shakespeare's plays are a great example of this sense of excellence. I've grown over the years to prefer Shakespeare to almost any other playwright, because no matter how many times I see *Macbeth* or *Hamlet,* the play is produced differently from the last time and the differences are part of why the production is good. Even the same producer and the same director and the same actors create a different *Twelfth Night* each time. That's the unique uses brand of excellence.

What kind of evidence is sought in a uniform impact assessment? Very sensitive instruments are specifically designed to pick up progress in a particular direction: progress in achieving the goal. Is this kind of evidence objective? Let's consider the role of subjective judgment and expertise in uniform impact assessment. A lot of judgment is used to design instruments that are valid and reliable enough to detect small differences in learning, the difference between a B and a B+, let's say. The subjective judgment embedded in these assessment instruments includes many somewhat arbitrary decisions about what particular performances can be trusted to stand in for the larger ability and about why that larger ability is worth attention.

One difference between the assessment of unique uses and uniform impacts is that the act of judgment is much more on the outside with unique uses. Although both types of assessment require expertise and subjective judgment, what judges have done has been buried underneath the fact of the tests in uniform impact. The test does not foreground the decisions that led to choices of features of this test or the expenditures making sure that the test does indeed measure what faculty expect it to measure.

In unique uses, on the other hand, students are assessed one at a time. The people who place a value on the learning of each student must be "connoisseurs," to use

Elliott Eisner's phrase. The external grader at Brown University whom I mentioned, for example, had been grading exams for years for many different teachers at Brown, all of whom taught different sections of the same literature course. When she said a paper was a B paper, there was a lot of expertise to give some credence to her judgment. She was a connoisseur.

To do a unique uses evaluation, we usually need a particular sort of connoisseur. We may be interested only in outcomes that relate somehow to a literature course, for example. But even within those bounds of novels and poetry and falling in love with words and understanding grammar, the connoisseur has a wide range of judgments to make, comparing apples and oranges.

How are the two perspectives on evaluation different when it comes to communicating findings in a convincing way? Some people assume that uniform impact is more credible, because decision makers only want numbers. Well, yes and no. About twenty years ago, Empire State College had a vice president for evaluation named Ernie Palola. I was visiting Empire after its evaluation shop had been in operation for several years. Ernie pointed out a format for reporting evaluations of which they were very proud. On top of a single heavy sheet of paper was a frequently asked evaluative question about this new college. Underneath was the answer to that question, usually in the form of a number and a table and a couple paragraphs of explanation. Each page was a self-mailer, so if somebody would mail or phone in that particular question, this sheet of paper was folded, stapled, and mailed to the inquirer. The report was brief, quantitative, and to the point.

Although Ernie was very proud of this way of communicating evaluative data to the public, he said wryly, "The paradox for us is that our most popular report, even now, is the first one that this office issued. It's about forty pages long, it has no pictures, it has no numbers, it's solid text." As I recall, this popular report was entitled something like "Ten Out of Thirty." Written after Empire State's first year of operation, it consisted of long narratives about several of Empire State's first students. Each chapter told a story of the encounter by the student with the institution, what the student did, and how well it seemed to work. Empire State College: one student at a time.

The stories added up to a story of a college, bigger than the stories of the individual students. As has been often observed, narrative is a very powerful way of teaching and a very powerful way of learning. Those stories were a great way to understand what this very strange institution was about and how well it was doing. I can't imagine numbers accomplishing this level of explanation and understanding, because numbers alone assume an unspoken context: how much or how many of some quantity that evaluator and reader both understand. With Empire State, there was no shared, vivid understanding. The stories helped supply that context. Without such shared context, the number may not be nearly as informative or decisive as the evaluator thinks it will be.

III. What the Good News Misses

The third thing missed by my straw man of evaluations that rely solely on outcomes assessment has to do with the obsession with good news. The false analogy between assessment and evaluation, on the one hand, and grading on the other, leads us too

often to design evaluations that focus on finding good news. That perspective, obviously, misses important stuff.

The obvious gap is that you need to detect problems before you can fix them. This is more than a cognitive issue. Uri Treisman once remarked, "Our problems are our most important assets." What he meant was that energy and resources flow to important problems. The more urgent and well-documented the problem, the more resources can flow to its solution. Not everyone realizes that problems can be assets. Some faculty members, for example, avoid using items that focus on worrisome issues because they don't want to look bad.

But if you think about it in Treisman's way of resources flowing to problems, imagine that you want to improve something about your program. Don't you need to be able to document that it's not working well in order to make the case that you're going to need more money or help? Now that's not to say that documenting a problem automatically leads to money, but it does mean that you're going to have an easier time crafting your request for more resources if you know more about what's going wrong. As a long-time FIPSE program officer, I can attest that we were much more responsive to proposals that began by graphically documenting a real problem for learners. Although there also had to be an opportunity to solve the problem, identifying the problem was crucial.

But there's a deeper sense of "looking for bad news" that I'd like to explore. I'll begin with the project I mentioned before about chat rooms, ENFI, Educational Networking for Interaction. Visualize a scene: In a classroom you see a circle of computers with big monitors. Students and a faculty member are sitting behind computers, not talking to one another, all typing. The dialogue of the class is appearing and scrolling up the screen.

ENFI provided a genre of dialogue that was midway between informal oral discourse and the formal written academic discourse that the students were trying to learn. This mid-level written conversation provided a very different ground and a different set of instructional possibilities for the faculty member. It was an exciting new idea at a time in the mid-1980s when the term "chat room" was not yet widely known.

Trent Batson, who had invented this approach, had asked the Annenberg/CPB Projects, where I worked, for money for a large-scale evaluation of this approach to teaching. He had assembled a team of faculty members from seven colleges and universities. When the Annenberg/CPB Projects funded the ENFI project, I, as the monitor of the grant, attended the first meeting of the faculty after their courses had gotten under way. It was about two months into the first semester, and the discussion among these faculty had been going on, as I recall, for about an hour and a half, maybe two hours. At that point, Laurie George, an English faculty member at the New York Institute of Technology, turned to her colleague Marshall Kremers and said rather quietly, "Marshall, you should tell your story." He said equally quietly that he didn't want to. She elbowed him a little bit and said, "No, you really should talk about this, it's very important." So he reluctantly began.

Kremers said that on the second or third day of class, the students in their writing had suddenly just erupted in obscenities and profanities that filled up everyone's screens. The professor became just one line of text that kept getting pushed off the screen by the flood of obscenities coming onto the

screen. Kremers kept typing "Let's get back on the subject" or "Won't you quiet down?" but the flood of student writing always pushed his words off the screen. Although he thought about pulling the lectern out from the corner and pounding on it, he decided, "No, this is an experiment; I've got to stick with the paradigm."

So Kremers walked out on his class. He came back later, either in the same class hour or the next class meeting, but it happened again: They blew him out of the classroom. It happened a third time. The fourth time, he told us, he managed to crush the rebellion. I don't think I've ever seen a faculty member looking more ashamed or more guilty over something that had happened in his classroom. He concluded by saying, "I don't know what I did wrong." And there was a long silence. And then somebody else in the room said, "Well, you know, something like that happened to me." Someone else added, "Yes, yes, something like that happened to me, too." It turned out about a quarter of the people in the room had had an experience something like that.

Diane Thompson, an English faculty member at Northern Virginia Community College, said, "Yes, something like that happened to me, too. But this is the *third* semester I've been teaching in this kind of environment. One of the things that I've learned is that we rather glibly say that these are 'empowering' technologies, but we haven't really thought about what 'empowering' means. Think about the French Revolution! Think about what happened when those people got a little bit of power. They started breaking windows and doing some pretty nasty things testing their power.

"But this is not all bad news. If you want to run a successful composition course, the really important thing is to have energy flowing into writing. And that's what you've got there, Marshall," she said. "The challenge here is not to crush the rebellion; it's to channel the energy!"

Well, all of a sudden everybody was talking about how to channel the energy. Meanwhile, I was sitting there thinking that I'd seen something like this before, at Evergreen. In fact, it happened pretty frequently, because Evergreen was unlike other teaching environments that most faculty had experienced. Faculty coming to Evergreen often blamed themselves for something that went wrong, something that actually happened pretty frequently, although they didn't know that because they were new to the institution.

But there were some differences between Evergreen and the situation in which Kremers found himself. First of all, Evergreen faculty always taught in teams, new faculty members being teamed with experienced faculty members. Experienced faculty would counsel a newcomer, "This is the kind of thing that happens at Evergreen. You may have done something particular to pull the trigger, but this kind of thing goes wrong easily at Evergreen. It's not a problem that can be easily eliminated or avoided. You can, however, build on our past experience. You might try this; you might try that."

That sort of conversation happened a lot at Evergreen. But Marshall Kremers did not teach in a team. If he hadn't been part of our evaluation team and able to learn with us, he might well have simply stopped using ENFI.

A second difference from Evergreen that also put Kremers at risk was that he was dealing with new technology. Because technology and its uses change every year, there isn't much chance to accumulate a history about what has been going on, the

way that Evergreen's veteran faculty understood the dilemmas posed for faculty.

I think often about the hair's breath — if Laurie George hadn't been there to say, twice, to Marshall, "You really ought to tell your story" — whether this experience would have come out at all. But she did prompt him to share his story, and I'm told that he has written a couple of valuable articles about it since then.

If we taught people to fly the way that we teach them to use most educational innovations, we would say to the not-yet-pilots, "Look. This is an airplane. It's really great for going all sorts of places. You could go to Portales, New Mexico; you could go to Paris; you could go almost anywhere you want. Now why don't you step into the cabin with me, and we'll take off. We'll fly around a little bit, and we'll land back here again. And then, I'm going to hand you the keys to the airplane, and if you want to go to Paris, it's east of here. This button on the control panel is the radio, and if you need a help line just push it, because we usually have somebody on duty, and hopefully they can help you if you run into trouble between here and Paris!" That's how we teach most faculty to use technology in teaching in their disciplines. We sell them on the technology and teach the rudiments, but we don't prepare them for problems they might encounter as part of the teaching activity. I define that as a career risk.

We ought to give faculty practice in "simulators," for want of a better word, that enable them to get into and then out of trouble in situations that are actually safe. One familiar example of a simulator is a teaching case study that is discussed by a seminar of faculty, but I don't know of any teaching case studies that spring from a technology-related problem like the one that hit Marshall Kremers. And I suspect there aren't very many that have to do with really innovative approaches to teaching generally; the ones I've seen deal with classic problems, not emerging ones. The use of simulators is awfully important because, number one, faculty members need to have a reasonably safe experience, safe to their careers, especially if they're junior faculty. It's very traumatic in technology. Junior faculty members are often advised not to have anything to do with technology until after they've gotten tenure, which is not exactly the way for a university or a college to make fast progress.

Now I can make my real point, about the good news that can be hidden in bad news. Remember that first observation that Diane Thompson made about the French Revolution and about empowerment. I've never thought about empowerment the same way since that day. Diane's observation about the dark side of empowerment gave me a richer, more useful way of understanding a whole range of phenomena. We gain a fuller and richer understanding of the strengths of what we are doing by looking at the problems that it causes squarely in the eye.

Here, too, my experience at Evergreen was helpful. I decided what core practices and goals to evaluate at Evergreen by first asking what problems the College couldn't definitively solve. Those dilemmas were the flip side of its strengths. It couldn't solve such problems completely without abandoning the corresponding strengths, so the problems remained unsolved. For example, a perennial problem at Evergreen was the student complaint of an insufficient choice of courses. That stubborn problem helped point my attention as an evaluator to Evergreen's practice of faculty teaching only one course at a time, sometimes for a full academic year, as part of a team. By deploying

its effort that way Evergreen was able to do many valuable things — it made narrative evaluations much more feasible, for example, and gave faculty and students the kind of flexibility I mentioned earlier. But one price was that the College could offer only a tiny fraction of the courses that a college its size would ordinarily teach. That problem was insoluble unless the College abandoned one of its core strengths. That's why an important part of my evaluation was then targeted on these full-time teaching and learning practices, because the insoluble problem had attracted my attention.

So dilemmas and core strengths are often the flip sides of the same practices. The more stubborn the problem, the more important is the underlying goal or strategy for the institution over the long haul.

Any program offers a wide range of practices and values. Which ones should an evaluator study? You can do worse than first looking for insoluble problems, and then using them to identify the most important, long-term goals and values.

Let's apply this kind of thinking to faculty development and new technology. I have a proposal to make. It comes in four parts.

1. Research to Identify Dilemmas

The first part is that I would urge faculty to do more research aimed at discovering the dark side of the force. Pick a new instructional situation, teaching courses on the Web, for example. Get people together who have had a little bit of experience with such teaching. Reassure them, "This is not going to get out; it's not going to destroy your career; it's just within this room. Now identify some of the most embarrassing things that have happened to you as a result of the thing you've tried to do with technology, or worrisome things, things that really frayed your nerves or whatever. It's probably something that never happened to anybody but you. That's OK. We want to share the really bad stuff, though." And then we'll wait and see whether other people say, "You know, something like that happened to me." Because we're going to be looking for the patterns, not necessarily universal patterns. Remember that what happened to Kremers only happened to a quarter of the people in the room. But if you've got ten or fifteen people there, things that happened to two or three people would be, I think, quite enough to be significant.

This important scholarship is something that many faculty members and institutions ought to do, because there are so many variations in what we do and thus so many dilemmas to discover. Because this research is time-consuming, no institution is going to be able to do it across the board. There is, therefore, plenty of room for lots of people to do this kind of research.

2. Develop "Simulators"

Second, based on discovered dilemmas, we then need to develop simulators — teaching case studies, role-plays, videotape triggers for discussions, computer simulations. Although I don't know what they all might look like, they would have in common their ability to enable faculty, teaching assistants, and adjunct faculty to encounter these kinds of situations in a safe setting where they can try out different sorts of responses. Many of these simulators will involve group discussion.

If you've never used a case study before, don't underestimate a case study by just reading it. Case studies are often not fascinating reading. After describing a problem, they stop. The case study itself is like the grain of sand in the oyster. The value is not in what you learn by reading the case. It's

Great outcomes might be achieved despite the tools rather than because of them; that's just one of many reasons why evaluations need to attend to means, not just ends.

the pearl that develops as people say, "Here is why I think the problem occurred and what I would do about it."

For example, I've been in other discussions about the kind of anarchy that Marshall Kremers discovered, and not everyone takes off from where Diane Thompson did, about empowerment. Other folks have different kinds of analyses about why Kremers's problem happened, and thus different ways of responding to it. For example, some might say that this kind of problem happens frequently in groups. Or other participants might point out that chat rooms can be fundamentally, subtly annoying because of the difficulty in timing your comments, so some kind of explosion is likely.

Each different analysis suggests a different set of indicators to anticipate, and different responses when trouble begins to develop. Because of the variety of possible analyses, I favor relatively unstructured simulators that give participants more freedom to suggest a variety of analyses of the problem.

3. Shedding Light on the Core Ideas

The third step is to brainstorm about the dilemmas and ask what strengths they reveal by their intransigence. Each dilemma can reflect the underside of a goal or strength, just as the Kremers anarchy reflects a richer view of an empowered student. After using such a simulator, the participants all can reflect: "What light does this shed on the larger situation? How does this change our ideas about the nature of what we're trying to do?" These kinds of role-plays and simulations can provide a setting for developing richer, more balanced, and nuanced insights into values and activities that are most important for the education of students.

4. Using Simulators for Faculty Development on a National or International Scale

Finally, we ought to make these kinds of simulators more widely available. A simulator developed for geography at a community college in Alaska may well have relevance to an elite selective private university. The biggest surprise in my visits to many institutions in this country and abroad is that while faculty members differ in the specifics of what they teach and learn, the dilemmas that they face are comparatively universal, across disciplinary lines, types of institutions, even national boundaries and language barriers.

For example, Kremers's experience with anarchy in a chat room can appear wherever chat rooms are used, which is in lots of fields and lots of settings. A teaching case study that had transcripts of how students exploded in a chat room environment could even be translated into other languages and be used appropriately in many countries around the world. Case studies developed in the United Kingdom could be employed in the United States.

How to get the simulators into wide use? There are many possibilities. For example, The TLT Group, of which I am a part, could be helpful in offering workshops around the world based on your simulators, face-to-face or online. I'm hoping we can collect simulators developed in many places and make the whole collection available internationally. Disciplinary associations could perform the same dissemination

function within their fields.

I think faculty could write and get funded proposals to create and disseminate simulators. Faculty could go in different directions and approach different funders to get support for doing simulators in their arena.

IV. Closing Remarks

My straw man — basing evaluation on the assessment of the average outcomes while looking for good news — is not a bad thing, but it's a radically incomplete way of evaluating academic programs.

First, studying strategy-in-use, not just outcomes, is really important. We must examine what people are actually doing to achieve the outcome. The Flashlight Project's tools, for example, prompt faculty to use data about strategy-in-use as a part of the story about why outcomes might or might not be changing. Look at people's satisfaction with the tools that are in hand when used for that strategy and that goal. Great outcomes might be achieved despite the tools rather than because of them; that's just one of many reasons why evaluations need to attend to means, not just ends.

Second, attend to unique uses, not just uniform impacts. Today's innovations, especially those using technology, tend to be empowering. Like the library, they increase the role of divergent learning: learning that is different for each learner. If we fail to use unique uses assessments and evaluations, we blind ourselves to a whole class of benefits and problems.

Third, look for bad news as well as good news, particularly because the worst pieces of news, the dilemmas, often are the flip side of what's most important about a program and shed some real light on the program's strengths. By developing simulators that help people cope with problems that cannot be definitely eliminated, you can protect the careers of the people who are working in your institution. And if you can help prepare them to deal with this bad stuff, they're much more likely to help their students learn. ●

References

Beeman, William O., Kenneth T. Anderson, Gail Bader, James Larkin, Anne P. McClard, Patrick McQuillan, and Mark Shields. (1988). *Intermedia: A Case Study of Innovation in Higher Education.* Providence, RI: Institute for Research in Information and Scholarship, Brown University.

Bruce, Bertram, Joy Peyton, and Trent Batson, eds. (1993). *Network-Based Classrooms: Promises and Realities.* New York: Cambridge University Press.

Chickering, Arthur, and Stephen C. Ehrmann. (October 1996). "Implementing the Seven Principles: Technology as Lever." *AAHE Bulletin,* 3-6.

Chickering, Arthur, and Zelda Gamson. (March/April 1987). "Seven Principles of Good Practice in Undergraduate Education." *AAHE Bulletin,* 3-7.

Ehrmann, Stephen C. (March/April 1995). "Asking the Right Questions: What Does Research Tell Us About Technology and Higher Learning?" *Change,* 20-27. Also available on the World Wide Web at <www.tltgroup.org>.

——— . "How (Not) to Evaluate a Grant-Funded Technology Project." (1998). In *The Flashlight Evaluation Handbook,* edited by Stephen C. Ehrmann and Robin Zúñiga. Washington, DC: The TLT Group. Also in draft form on the World Wide Web at <www.tltgroup.org>.

.....

Stephen C. Ehrmann is vice president of The TLT Group, the teaching, learning, and technology affiliate of AAHE, and director of the Flashlight Project.

Ehrmann's path to this conference began with his experiences as director of educational research and assistance at the Evergreen State College, and then as program officer with the Fund for the Improvement of Postsecondary Education. In neither venue was it enough to look just at outcomes if interested in understanding or improving an innovative practice. Evergreen also helped him see how an innovation's insoluble problems often reveal insights into its core values.

In 1985, he became a program officer, later senior program officer, with the Annenberg/CPB Projects. During 1990-93, he also served in that role with the Annenberg/CPB Math and Science Projects, to support national reform in K-12 math-science instruction. At Annenberg/CPB, he was able to support a number of evaluation-intensive projects that shed important light on the variety of outcomes that can arise from the use of empowering technologies. There he also created the Flashlight Project, to develop evaluation tools and services for studying the consequences of technology in education. In 1997, Ehrmann moved to AAHE. In 1998, he helped create The TLT Group as an independent, nonprofit company. His varied experiences with The TLT Group as consultant, speaker, and evaluator, along with his work on distance-learning strategies, have helped him further develop his thinking about the nature of evaluation.

Assessment of Powerful Pedagogies: Classroom, Campus, and Beyond

by Jean MacGregor

Let's begin with a quick assessment of who is here. Who of you consider yourselves "novices" at assessment, perhaps at this conference for the first time, just getting your feet wet, or even just furtively eyeing the water? [More than half the audience.] Who of you would term yourselves "intermediates," individuals who have waded into the water and are taking your first strokes? [About a third of the audience.] And who of you are the "experienced" in the room — people who have been swimming in this water for some time? [Less than 10% of the audience.]

Let's also see what roles we play in higher education settings: Who here is a college student? A faculty member in the classroom? Individuals with expertise in assessment or evaluation who are working with faculty members as mentors, coaches, or coinvestigators? Faculty members in K-12 education or postsecondary education? Administrators with responsibility for assessment? Program officers or staff of public or private foundations with interests in assessing student learning in the grants they make? It probably goes without saying that individuals in every one of these roles should be involved in assessment efforts, and that assessment is the healthiest on those campuses where the process engages a diversity of individuals in common inquiry.

I have to admit some embarrassment, seeing myself named as an assessment expert in the conference program. In the categories above, I feel most comfortable being termed one of those intermediate individuals getting into the assessment water. I would like also to identify myself in my role list first as a faculty member coming to assessment in my own classrooms with my students, and second in that coach and colleague category, bringing assessment

ideas to my teaching colleagues as they undertake new work — new moves in their teaching — and also probing for how they think about outcomes for students and ways to realize them.

As my remarks this afternoon will illuminate, my insights about assessment in the context of "powerful pedagogies" have emerged from working with some communities of educational pioneers in Washington State through the Washington Center for Undergraduate Education, a public service initiative at the Evergreen State College.

The Washington Center is a partnership of campuses, both two- and four-year, working in a grassroots fashion on issues of curriculum development, faculty development, and assessment. We also support academic success for students of color with several institutional assessment and capacity-building projects. The Washington Center was founded thirteen years ago out of an exciting collaboration between Evergreen and Seattle Central Community College, and that partnership of two campuses has grown over the years to forty-six campuses — nearly all the public and private institutions in Washington.

In several of our projects, my role has been a bit Perle Mesta (convening conversations), a bit Johnny Appleseed (traveling around picking up and planting seeds of good ideas), and a bit Saul Alinsky (organizing on behalf of institutional change to support innovation and reform efforts). Our assumption at the Washington Center is that within any one state or region, there are great reservoirs of talent and interest in curriculum and teaching improvement, but there need to be vehicles to share that talent and to build on it.

My introduction to this strand of the conference is divided into four parts:

1. A brief overview of what we are calling "powerful pedagogies" and the ways assessment of them appears in the conference.
2. Some frameworks or concept maps for navigating the assessment territory.
3. Reflections on ways that assessment emerged in two grassroots curriculum reform efforts in Washington State.
4. Some thoughts on what kind of assessment efforts are required to support these emerging pedagogies.

The first two parts of these remarks are more explanatory, the second two more exploratory. I would like to emphasize that this work of new pedagogies is so diverse, both the development of new teaching approaches as well as very creative assessment approaches, that any one of us just has a few jigsaw puzzle pieces of a picture that I think — I hope — will emerge more tangibly in the coming years.

Powerful Pedagogies

In the last two decades or so, we have seen gathering streams of exciting work deepening our understanding of the human learning process. Both Peter Ewell and Ted Marchese have made masterful attempts at summarizing and distilling these streams for us at this forum, in *Change* magazine, and in the *AAHE Bulletin*. I won't march you through all of their points here, but would like to nod to the rapidly expanding bodies of literature on human learning that Ted described in detail last year.[1]

The field of developmental psychology has been expanding steadily in its views on the intertwined patterns of intellectual and

[1] "The New Conversation About Learning," in *Assessing Impact: Evidence and Action,* 79-95 (Washington, DC: AAHE, 1997).

ethical development of students. More recently, cognitive psychologists have been positing new, complex ideas about cognitive development and learners' ability to construct knowledge from infancy right through their lifetimes. The field of neuroscience has been growing exponentially in recent years, providing information on how, physiologically, the brain learns. At the same time, new studies of learning are emerging from anthropology. Some significant ethnographic research is occurring in settings other than schools, primarily apprentice programs. In addition, there are ethnographic studies on workplace learning and the richness of workplaces as contexts for learning. Adding to the storehouse on human learning is even the field of archaeology, through research on prehistoric brain development.

Working in parallel and also drawing on some of these literatures is higher education research. Several studies have emerged in recent years on student learning in college, much of it coming out of Western Europe, on what and how students learn. Some of this literature distinguishes surface learning from deep learning. Surface learning is that which is taken in and memorized in superficial ways only to be discarded and forgotten; deep learning is that which is so firmly rooted that students can see its applications and can draw on it to use in different and new settings. As one of the students in our learning communities in Washington said, this latter kind of meaningful, lasting learning is "real learning" as opposed to "just learning."

These strands of work give us new conceptions and new vocabularies for expanding our own mental models for how powerful learning occurs — or doesn't — or for simply affirming what we have sensed, observed, and practiced in our own teaching. And although no one has attempted the grand unifying theory of learning for the 1990s, there is considerable crossover of ideas and linkages between the findings in these various fields.

At the same time, active communities of practice on our campuses are engaged in a variety of efforts to improve specific courses, bodies of coursework, and curricular and cocurricular experiences for students. Many of these efforts draw specifically on the research literature I've just mentioned, and others are rooted in reformers' own experiences and intuitions about what is effective for student engagement and learning. These improvement efforts usually appear on lists of "alternative pedagogies" or "powerful pedagogies," if you will. They include:

- Collaborative and cooperative learning involving ways of structuring learning situations so that small groups of students construct meanings together or create a product of some sort; also situations in which students act as mentors or coaches to other students.
- Active and interactive learning strategies having to do with writing, and often with technology.
- Problem-centered learning and case-centered learning in which an open-ended, rich, and puzzling problem challenges students, often working collaboratively, to take apart the problem, marshal information to work through the problem, and offer their best attempt at an analysis or a solution.
- Service-learning and civic learning and other forms of experiential learning that link theory and practice and put students directly in touch with local communities and community issues.
- Interdisciplinary courses and learning communities in which the curriculum is

Of course, putting forward these new strategies begs the question "Why?" If we are immersing students in these kinds of contexts, to what ends? The phrase of the day is that "It's about learning." Yes, it is about learning, but our intentions are quite a bit more complex than that.

literally re-formed around interdisciplinary ideas in order to engage students in more holistic explorations of boundary-crossing topics or ideas.

- Curricular and cocurricular interventions that link academic work with student life activities that increase the chances of student success in college.
- Capstone experiences, a traditional staple of senior-level offerings in the major of liberal arts colleges, with exciting variations such as with internships, applications projects, and interdisciplinary research projects — a powerful end-of-college-career assessment occasion.
- Assessment as learning, an approach not often on lists of "alternative pedagogies" but which absolutely should be if we take assessment to mean a process of embedding assessment very explicitly in any teaching setting; that is, students are asked to recognize what they bring to the learning experience, the outcomes or areas of competence are made clear, teachers are explicit with students about learning strategies to build competence, give them chances to demonstrate that competence, and give them feedback over and over.

These approaches are not distinct; you can probably think of several projects on your own campuses that incorporate several of these pedagogies simultaneously. That is what makes these emerging approaches so interesting — as well as challenging to assess.

Looking for a pattern in these approaches, I think it is fairly clear that they all move from a mode of college teaching and learning that is content-driven and delivery-oriented to one that is more student-oriented and learning-oriented. Classrooms are less centered in teacher performance and more centered in expectations of student performance; that is, classes are less performance settings for teachers and more practice and performance settings for students. And that means that faculty are designing conditions for student learning but refusing to bear the whole responsibility for the class' agenda and success. A great deal more responsibility is placed on the students. In this conference, you will see sessions built around assessment of student experience and learning in these arenas of alternative pedagogies.

Of course, putting forward these new strategies begs the question "Why?" If we are immersing students in these kinds of contexts, to what ends? The phrase of the day is that "It's about learning." Yes, it is about learning, but our intentions are quite a bit more complex than that. Although many different typologies exist for these goals for students, these goals are emerging for many of us at deeper levels as we consider ways to see them demonstrated in student work. Here is my list of intentions or ends that these pedagogies imply:

- knowledge that students retain over time; learning that has real meaning; learning that students can apply in new contexts
- thinking, reasoning, and problem-solving skills in specific contexts
- "information literacy" skills
- communication skills, especially across significant differences
- collaborative skills and abilities to work in teams

- metacognitive and self-reflective skills — the ability to look at one's own learning, to build the capacity to think about learning, and to assess one's strengths and weaknesses
- competence in a field of study or a professional concentration
- aesthetic perspectives and values abilities, sensibilities and values for living and contributing in a pluralistic society, a participatory democracy, an ever-changing, complex world.

Educators in higher education have been talking about these outcomes for years. They are the kinds of lists that we struggle with and negotiate about in general-education committees. Yet, I sense that two trends are emerging as we talk about these outcomes. One is that we talk less about the domains of knowledge for college graduates and more about abilities and sensibilities that we want to foster. As Buddy Karelis at FIPSE (Fund for the Improvement of Postsecondary Education) says, it's no longer just about having the right array of cans in your shopping cart when you get to the check-out line in the grocery store; it's more about what you understand about food and how you think about putting it all together to make a healthy meal.

Second, if we commit ourselves to outcomes like these in all their complexity, then we need to move to explore much more carefully ways in which these outcomes develop and ways to immerse students in learning settings that elicit these outcomes. That's the challenge before us.

Several sessions in the conference speak to assessment for these outcomes — most particularly at the level of the individual course but also at the level of the program. Some presenters have developed assessment tools that faculty members can use to assess for a certain outcome or can adapt for their own purposes, while other presenters have worked with faculty directly to invent and embed assessments in their existing courses and to elicit evidence about the outcomes. Powerful assessment tools can be a useful, powerful avenue for starting conversations about our goals for student learning and about alternative teaching approaches.

Navigating the Assessment Territory

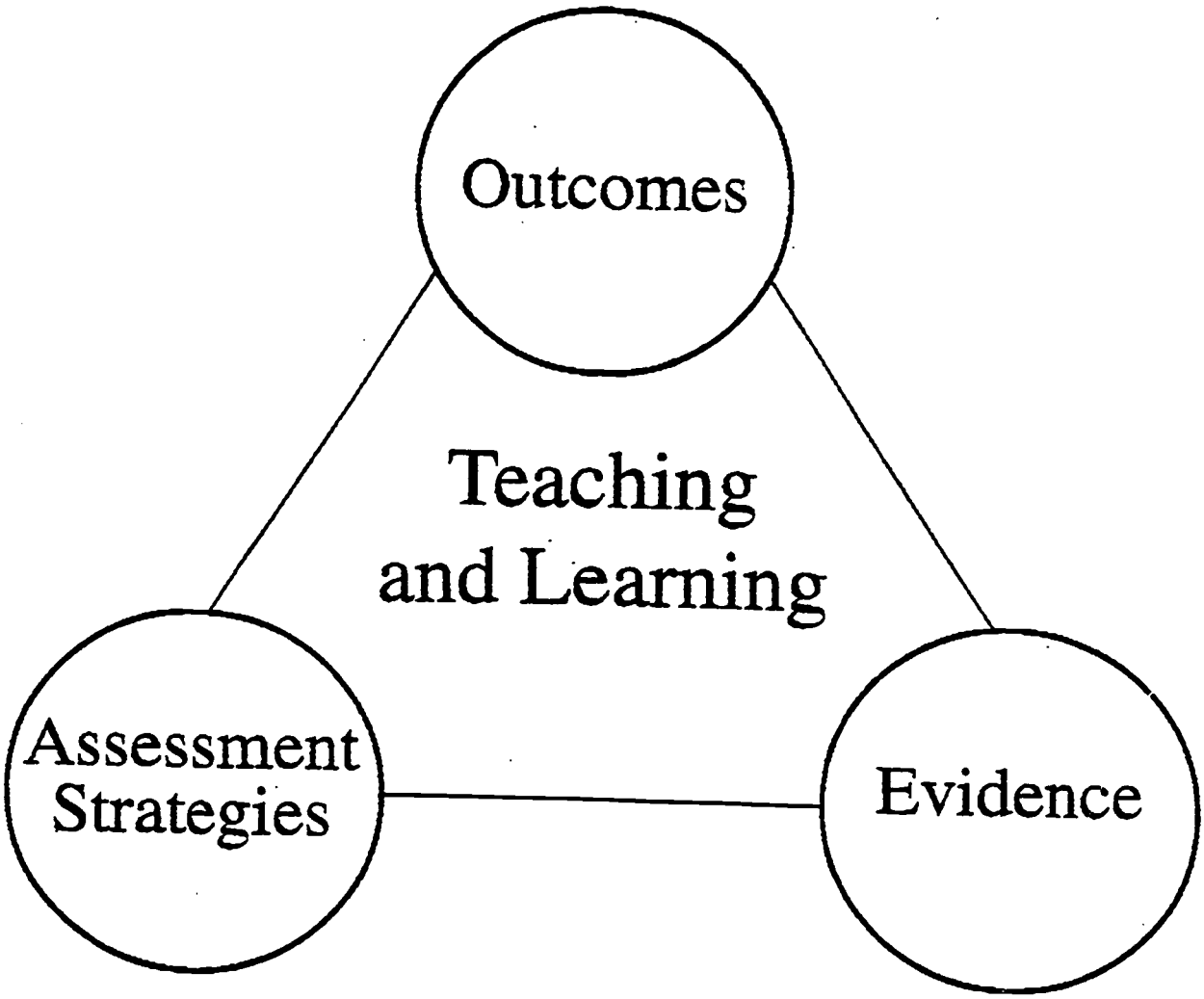
Assessment as an emerging practice in higher education is complicated to enter at first, because it can occur on so many levels and because the term is used broadly to refer to so many different kinds of specific strategies. Two constructs have been useful to me as I have waded into the water; I hope they'll be helpful particularly to those of you who are just getting your feet wet.

The first is to think of assessment as a process rather than a particular technique or an instrument. Figure 1 *(on the next page)* is my three-legged-table scheme of assessment. For this table to stand up, we need at least these three legs. On the table, for sake of this conversation, could be one of those powerful pedagogies. Or, we could put on the tabletop a course or a whole general-education curriculum. Starting at the top of the table, we have goals or intentions for that teaching approach that have to do with student learning outcomes. This is the ideal, in our imaginations, of what success would look like in student work or student performance. What is on the table — our teaching and learning strategies — should ideally resonate with those intentions.

We choose and carry out strategies for gathering information about whether and how students are meeting those outcomes (on the lower left side). Then we make choices or interpretations about that information that we communicate to others as evidence — evidence of student learning or

Figure 1

Assessment as a Continuous Process



evidence about the student experience. This evidence is the real, on-the-ground results that we communicate to students, colleagues, or external bodies, about how close we came (or didn't) to meeting our intentions.

This triangle oversimplifies, of course, a complex process of choice-making at every step of the way, but as you move through the conference, it might be useful for recognizing the points at which various individuals and campuses are working. When I am learning about an assessment practice, I like to ask, "What are the intentions or outcomes for the educational program? What are the stated (or often unstated) assumptions about what success would look like for students or for student learning? What kinds of teaching/learning situations resonate with the hoped-for outcomes? What kinds of assessment strategies are occurring? Do these strategies resonate with the learning goals? What kinds of information or evidence resulted about student learning or the student experience? What sense is being made of that information? Who knows about the information? Did having the evidence make any difference? Did communicating the evidence make any difference?"

For individuals with extensive assessment experience, this little visual may seem like a firm grasp of the obvious, but in my travels among faculty who have devoted much of their teaching careers to a delivery-and-explication model of teaching, this construct is very foreign. For many, assessment is not a language with which they are familiar. Others seem to think assessment is only about evidence-gathering strategies or the imposition of instruments — not something they can design and control. They often don't recognize that this model is quite useful for framing thinking about course design and teaching.

A second construct that I've found useful in thinking about assessment sketches out different purposes for assessment. *(See Figure 2 on the next page.)*

If the assessment is about gathering evidence about student learning, who is this information for?

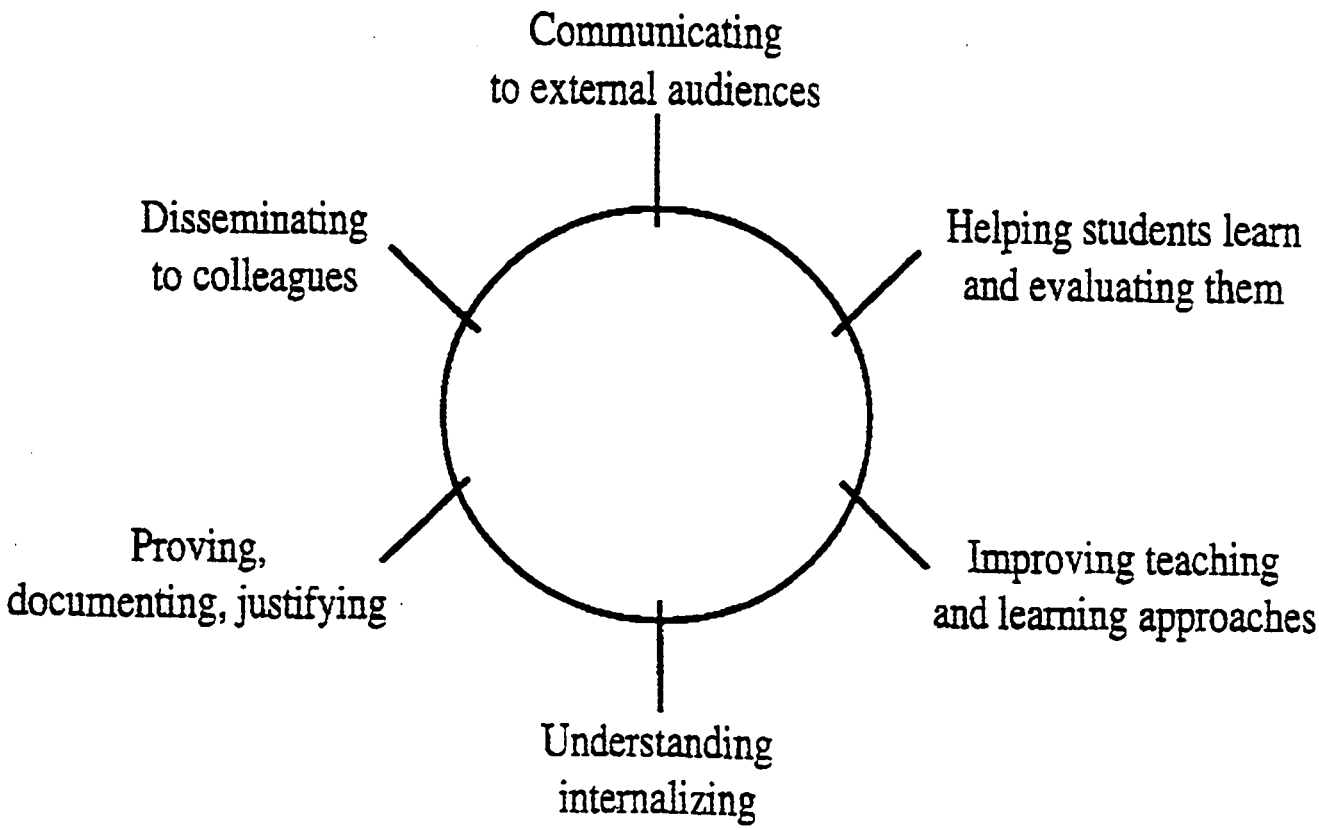
Let's start at the two o'clock point of this concept map and move clockwise. Assessment can be seen entirely at the two o'clock space as a process in an individual class, whereby I gather information about my students' learning, give them feedback on their learning, and evaluate them. Moving to the four o'clock point, assessment also can be seen as a process of gathering data about what students are learning and how they are responding to the teaching setting, in order to improve the course or program. Tom Angelo and Pat Cross have made a huge contribution to assessment practice with their compendium of Classroom Assessment Techniques (or CATs, as they're sometimes called): short, in-class, informal strategies for eliciting student feedback on their learning.

In the past decade, the most widely adopted new classroom strategy in the country probably has been the simple "minute paper" activity. At the end of class, the teacher asks students to write a sentence or two about the main ideas they have learned from the class and to ask a question about something that is still unclear. The one-minute paper and dozens of other simple information-gathering strategies can be enormously useful in giving teachers immediate feedback on what students are understanding and in including them as partners in the teaching/learning process. They also ask students to pay more attention to their learning in class.

Moving down to six o'clock, particularly important for teachers inventing new

Figure 2

Using Assessment Information



curricula and making new moves in their teaching, are opportunities for faculty self-reflection that goes deeper than simply evaluating students and making improvements to one's classes. I'll be talking a little more about this shortly. At the eight o'clock hour, another critically important use of assessment information is to document the effectiveness of an approach — especially if it's a new curriculum or program — and to prove whether the program is or isn't living up to its intentions or its claims. Still another application, at ten o'clock, is using that information to communicate to colleagues about approaches and results. Finally, there are the purposes of using assessment information with more external audiences: the institution-at-large, trustees, alumni, parents, the wider community, accreditation bodies, or funding agencies and organizations.

Once again, those of us who have been in the assessment pool for a while have internalized these purposes and levels and can easily see the distinctions between them. However, for faculty and staff new to assessment, there is understandable bewilderment, and it is no surprise that there are questions about assessment's purposes and audiences.

Assessment in the Context of Curriculum Reform

Moving to my own work in the arena of "powerful pedagogies," I want to describe two reform efforts with which I've been involved. One is a learning communities effort that actually propelled the creation of the Washington Center network thirteen years ago. The term "learning communities" is used widely to refer to a variety of efforts involving collaboration and community-building, but I am using the term here to refer to curricular approaches that link or cluster classes, often around interdisciplinary themes, and enroll common cohorts of students. The intentions for these course-linking or course-clustering approaches are multiple: student engagement and success through the creation of community and a holistic, interconnected learning experience; curriculum coherence, especially in fragmented general-education offerings; interdisciplinary curricula and the opportunity to organize coursework around compelling themes; and faculty revitalization — opportunities for faculty members to work collaboratively across disciplines and to share teaching approaches with a common cohort of students.

Learning community curricula are highly variable: They link courses from virtually every discipline. While most programs are geared to first-year learners and involve general-education courses, learning communities have been developed for underprepared students, for honors programs, and for study in the minor or major.

The teaching approaches used in learning communities are also variable, but generally they include a great deal of collaborative learning, integrative projects and assignments, self-assessment, and writing in the context of disciplines or interdisciplinary topics.

A few learning community examples are illustrative: "Revolutions and Reactions" integrates coursework in English composition, art history, and European history; "Chemath" links intermediate algebra and precollege chemistry for underprepared students; and "The Power of Place" is a team-taught program linking an American studies/humanities course on the American landscape with a freshman writing course.

The second reform effort with which I have been associated is reform calculus.

Our intention in the Washington Center was not to disseminate any one model of learning communities or any one reform calculus text but rather to create opportunities to explore potentially powerful reform ideas and at the same time to build networks of faculty in the state in the two- and four-year system.

This national reform effort grew out of a national conversation beginning in the late 1980s that finally said out loud that not only were students failing calculus all over the country but calculus was failing students in a host of ways. A group of reformers argued that calculus should and could be a vehicle for pumping students into advanced coursework in the sciences and calculus-requiring majors rather than filtering them out. Shortly thereafter, the National Science Foundation (NSF) funded several ambitious reform calculus textbook writing projects that centered around the following intentions: to reconceive and actually reform calculus courses to more successfully attract students to the math major and to other calculus-requiring majors; and to use a reformed calculus to spur deeper conversations about the entire math curriculum.

The components of reform calculus were: a "lean and lively" calculus that would prune back massive textbooks and focus on key calculus concepts; a pedagogical approach that would embrace multiple ways of learning calculus concepts — students would learn calculus not only through symbolic manipulation (what the math community refers to as "plugging and chugging" the numbers) but also through visual understanding and conceptual understanding; the use of electronic technology to solve calculus problems with both computer software and the new hand-held graphing calculators; and the setting of calculus in meaningful applications problems done in small-group settings. The idea here was to enable students to see calculus at work in real-world settings.

About twenty-five campuses in our Washington Center network became actively involved in experimenting with learning communities or with reform calculus. Both initiatives were incorporating many of the "powerful pedagogies" put forward at the start of this speech. However, they also were embedding — and this is key — these pedagogies in new ways of conceiving curricular content and structure. Energetic experimentation flourished, and continues to flourish, even though no external money was available to fund or release faculty to undertake this work. We did have some modest NSF money to fund workshops for the would-be calculus reformers and to distribute reform calculus curricula, but by and large all this reform work was voluntary and grassroots, and carried out on campuses without the infusion of external money.

Our intention in the Washington Center was not to disseminate any one model of learning communities or any one reform calculus text but rather to create opportunities to explore potentially powerful reform ideas and at the same time to build networks of faculty in the state in the two- and four-year system. Our strategy was to hold a series of retreats and conferences to put out menus of ideas that faculty could pick and choose from and adapt to their own purposes. These gatherings varied from small overnight meetings at church camps in the woods to substantial conferences of 300-400 participants in Seattle. Along with those gatherings, my colleagues and I made ourselves available to do site visits to campuses, to stop in and ask how it was going,

and to use that information to design further retreats, conferences, and newsletters. So, both reform efforts were big, ambitious, messy projects that played out differently on each campus.

Although the stories of faculty experiences in these programs are fascinating and countless, let us focus on how assessment played out in these efforts. With each initiative, we invited volunteers to serve on an assessment committee or working group, not to insist that this group conduct some sort of grand evaluation of all the reform work going on — it would have been an impossible task anyway, given the diversity of what was being tried — but simply to entertain the idea of assessment approaches and to begin to think of ways assessment could be used to further and deepen the reform efforts. Parallel conversations and efforts emerged with both the learning community pioneers and the reform calculus experimenters.

Referring back to the "Using Assessment Information" concept map (Figure 2), both assessment groups gravitated where you would expect them to — to the right side of the circle. Faculty members wanted to explore the connections among their new curriculum content, their goals for student learning, and appropriate strategies for assessing that learning.

The calculus group particularly wanted to clarify outcomes for the curricula they were adopting, because the different reform calculus texts were offering a multiplicity of emphases and directions to pursue. At one calculus retreat, we spent several hours brainstorming and then prioritizing our outcomes for reform calculus. It was fascinating. There wasn't perfect consensus, but everyone went home realizing the math department back on their campus needed to have the same conversation. And many did — especially in the context of asking, "If these are our outcomes for calculus, what are the implications for precalculus and for the other courses that come before and after?" So the reform curricula were actually forcing a needed dialogue about intentions and goals.

Further, the reform curricula, if they were to take seriously such "powerful pedagogies" as collaborative learning and writing-to-learn activities, were also pushing a conversation about assessment strategies. The effort stimulated not only conversation but also the active creation and gathering of strategies and approaches. The calculus group ended up creating a 600-page sourcebook on problem sets and test questions in order to better evaluate students, as well as a variety of classroom assessment techniques with which to assess student responses to reform calculus content and to new teaching approaches. The creation of this sourcebook effectively made the link between reform calculus curricula and a reformed pedagogy.

The learning community assessment group, though working across many more disciplines, was similarly interested in discussing goals and intentions for their learning community teaching and in developing appropriate assessment tools for evaluating students. Also, this group wanted to share ideas for classroom assessment strategies appropriate to collaborative learning settings that would give them information about the student experience in learning communities.

In parallel fashion to the calculus group, the learning communities group compiled a resource book of assessment approaches appropriate to learning community settings. One approach that particularly captured learning community teachers' interest was student self-evaluation — the process of

asking students to reflect in both formal and informal writing assignments on both the content and process of their learning experience. Faculty members saw student self-evaluation as especially promising because it serves multiple purposes simultaneously: It is powerful pedagogically in enabling students to describe and synthesize their learning in interdisciplinary contexts; it is useful for illuminating what students identify as important or problematic in learning community programs; and, it is a promising source of assessment information about student learning and the meaning students make of their learning experience.

Both groups expressed great appreciation for opportunities to come together to reflect and internalize on the new curriculum content, new pedagogical strategies with which they were experimenting, and new strategies for assessing student learning and gathering student feedback. Much of this collaborative reflection naturally took the form of storytelling about particular classroom situations.

This was right about the time that AAHE was developing its program in the use of teaching cases as a strategy for deepening conversations about teaching and learning. The learning community assessment group immediately saw the connection, and a dozen or so of its members became a case-writing group, shaping their stories into teaching cases about issues of learning community teaching as well as administrative implementation. As hoped, the casebook that resulted found its way back to the learning community–adopting campuses, where the cases were used in faculty-development workshops.

So, the assessment work that was of most priority to these reformers was building competence and confidence with new curriculum, new pedagogies, new approaches to evaluation, and new ways of gleaning student feedback. But when it came to assessment for purposes of proving or justifying their reform efforts (moving over to the other side of the circle) there was some resistance on the part of these faculty reformers.

This was in part an issue of limited time. Already busy, often intensely overcommitted, teachers taking on very exciting but very demanding new teaching projects wanted to focus on first things first: They wanted to explore the innovation itself and build their confidence. Their attention was on getting into the pool and learning to swim and on enabling the students to swim and not to get cold feet or drown. They were not ready yet to measure the speed of getting across the pool or to describe the elegance of student strokes to others.

There were additional issues. A prevailing perceived barrier to program assessment was faculty members' lack of evaluation expertise. Most of these experimenters were not social scientists and were very new to assessment concepts and practices. They felt daunted by the challenge of designing and carrying out comprehensive outcomes assessments. A third issue was the obvious tension about role — this is what my colleague in learning community work, Faith Gabelnick, refers to as the poet/critic paradox. "Here we are," she says, "encouraging faculty to be poets, to invent new curricula and ways of teaching. Can we or should we ask them to be critics at the same time of their own poetry?"

It's one thing to gather classroom assessment feedback to improve my teaching or to say that my innovation lives up to my own intentions for student learning, but it's quite another to say that the innovation in my classroom produces more and better learning than yours. Yet, eventually, just

these issues and just these comparisons are going to have to be put more squarely on the table and in fact are being put on the table as we speak, when it comes down to choices that institutions are having to make about resources: e.g., What level of class enrollment is most effective for a learning community program? Will a large class be broken into discussion sessions? Can we afford to undertake special training of teaching assistants so they can facilitate problem-centered learning in discussion sections or labs? Which introductory calculus text will be adopted department wide? What resources will be made available for field trips and field equipment?

Although some modest studies had been conducted of learning community impact on students, we turned a truly significant corner in learning community work when Vincent Tinto, of Syracuse University, and his team of graduate students proposed to do an intensive study of three learning community programs in the country, two of them in Washington State. With resources from the federally funded National Center for Teaching, Learning & Assessment at Penn State, Vince and his students carried out a sophisticated qualitative and quantitative study of the student experience in these programs, and they disseminated it widely. Subsequently, three doctoral students carried out dissertations on learning communities in Washington.

There is no question that the learning community effort nationally has been strengthened by these external researchers with the time, the formal role, and the resources to carry out detailed, credible studies.

Lessons From Washington State

So here are some lessons learned from supporting these two reform efforts and reflecting on the role assessment has played:

It's one thing to gather classroom assessment feedback to improve my teaching or to say that my innovation lives up to my own intentions for student learning, but it's quite another to say that the innovation in my classroom produces more and better learning than yours.

1. Powerful pedagogies cannot stand apart from discussions of curriculum content and structure. Outcomes conversations need to focus not just on skills and abilities but also on the "key content" of learning in courses or learning community programs. A common resistance to all these activities is, "Are we sacrificing coverage?" — which begs the question of asking what's really important to cover or what's really important for students to know.
2. There's no telling how or when faculty members will embrace an assessment framework as a way of thinking about designing, carrying out, and evaluating learning experiences. Some get it immediately; others find it opaque. I think a powerful way in the door is discussion about the test or other demonstrations of student learning: What do our tests or assignments imply about what we value in student learning, student knowing, and student abilities? What evidence do we draw from tests to evaluate students or to portray to others what students have learned?
3. The language of reform can be tricky because whether we use the term "reform" or "innovative" or "powerful pedagogies," we are, by implication, saying the rest needs reform, is obsolete, or is less than powerful. We need a

language to share and deepen our work, but we also need ways to talk about our innovations without creating divisiveness or marginalizing our colleagues by implication. Perhaps we should be calling these approaches "emerging pedagogies" or "promising pedagogies" at this point rather than "powerful" ones until we have more solid data about their impact.

4. Embracing new pedagogies takes time and a culture of permission to experiment. The current hype around the so-called "power" of certain approaches sometimes implies instant success, when in fact faculty using these approaches are often struggling, making missteps, and experiencing gains and losses in confidence on a weekly basis. Many studies of workplace learning point to the value of trial-and-error learning and the learning that comes from recognizing our mistakes and then really understanding and internalizing them.

   We need to create spaces that are safe enough for experimentation and failures. Pioneers and experimenters also need time and support to build competence and to internalize new ways of teaching. Conversations with like-minded or more experienced colleagues are invaluable. Team-teaching is almost priceless.
5. Third parties are critical for carrying out intensive, data-rich assessment work. Innovators truly benefit from alliances with evaluation professionals.
6. As important as the reforms themselves, and their value to students, are the communities of inquiry that can be created. In Washington State, the learning community leaders and practitioners and the reform calculus community are still gathering periodically to reflect on their work and share ideas.

Building Communities of Inquiry Around the "Emerging Pedagogies"

Moving from these lessons from Washington to the larger agenda of strengthening these new pedagogies and building solid bodies of effective practice, we have to take the concept of "communities of inquiry" to a much more sophisticated level.

First, we must bring our students more centrally and consistently into our communities of inquiry. As shown through the practice of classroom assessment, if we seriously ask students about what and how they are learning, and if we take what they tell us seriously as well, we discover that they are interested and even eager to give us feedback. These approaches not only give us important information with which to strengthen our teaching, when used regularly they can build important reflective capacities in students.

Another way students can be involved is by giving us feedback through instruments that go beyond the standard and often mind-deadening end-of-course course evaluations (which usually focus just on the performance of the teacher) to provide feedback on which elements of a teaching environment are working or not working. Steve Ehrmann's Flashlight Project is a fine example of a toolbox of instruments and questions that technology-oriented projects can use to elicit student responses on what's working. Some exciting work is emerging out of science reform efforts as well, through the development of instruments that tease out what pedagogical elements of a reformed course work or don't work for students.

Second, in our own institutions, we can create more formal, more extensive com-

munities of inquiry about student learning by linking curriculum reformers and pedagogical experimenters with expert social science researchers. At this conference, it is exciting to see any number of examples of projects in which individuals with evaluation or assessment expertise partnered up with faculty involved in the new pedagogies to carry out an assessment project.

For example, Portland State is launch ing a Classroom Research Resource Team to provide a forum as well as resources for faculty who wish to undertake research on student learning in their classrooms. A similar project is in place at the University of Wisconsin-Madison. A percentage of several large science and engineering curriculum reform grants has gone to an in-house evaluation staff, the LEAD Center. LEAD (Learning Through Evaluation, Adaptation, and Dissemination) is staffed by a team of expert quantitative and qualitative researchers who work with the faculty reformers, helping them design both in-class, formative assessments and also more summative evaluations of these projects. These two examples are real beacons of what every institution should and could undertake if we truly were to get intentional about "organizing for learning."

Finally, we must get more serious about expanding bodies of practice in these new pedagogies, both nationally and internationally. In any given year dozens, perhaps hundreds, of little assessment projects are under way in individual classrooms, which are valuable in their own right for advancing teachers' practices in these new pedagogies and for engaging teachers in thinking in new ways about assessment.

But at this juncture, all they are is just that: data points — often invisible data points. We need more than data points. We need information about results, about patterns, about trends. We need more synthesizers willing to assemble what we know about practice and what we know about results. We need not just bodies of practice but bodies of evidence.

In any given year dozens, perhaps hundreds, of little assessment projects are under way in individual classrooms, which are valuable in their own right for advancing teachers in thinking in new ways about assessment. But at this juncture, all they are is just that: data points — often invisible data points.

At this conference, Len Springer and Jim Cooper will be reporting on a research study conducted at the National Institute of Science Education that examined hundreds of studies of cooperative learning in college science, math, and engineering courses, and conducted a meta-analysis of some of those studies. The findings are pretty impressive regarding small group learning in this arena. And Susan Ganter, now at AAHE, has spent the past year at the National Science Foundation poring through all the reform calculus studies to distill out the patterns of results of eight or so years of reform calculus work.

And on the horizon we have the promising new Carnegie Teaching Academy, a partnership between the Carnegie Foundation for the Advancement of Teaching and AAHE. Its national fellowship program will bring together outstanding faculty to investigate and document their work on research projects on teaching and learning undertaken in their classrooms. Presumably, the emerging pedagogies will be prominent in that effort.

So I return to where these remarks began. It is exciting that we can look to the rapidly expanding and diverse communities

of inquiry and bodies of research related to human learning. In parallel, we have a responsibility to develop communities of inquiry and much larger research bases related to pedagogies and the most effective ways of fostering human learning. Disseminating information about the rationale and technique of various approaches has value, of course, but we need to move much more systematically to documenting our results, with our students, within our institutions, and more widely as bodies of practice grow. And that is where assessment in all its forms is so key. ●

.....

Jean MacGregor directs the National Learning Communities Dissemination Project (FIPSE-supported) at the Washington Center for Improving the Quality of Undergraduate Education at the Evergreen State College. The Washington Center, a public service initiative at Evergreen, is a grassroots partnership of forty-six colleges (both two- and four-year) in the state of Washington, working on various faculty, curriculum, and organizational development initiatives. In the past decade, the Center's work has revolved around the development of learning communities (course-linking strategies to foster interdisciplinary studies, integrative learning, and student involvement); math and science curriculum reform; and diversity (both cultural pluralism in the curriculum as well as academic success for students of color). Several of these projects have had strong assessment components growing out of faculty members' interests in strengthening and sustaining innovative efforts.

MacGregor has longstanding interests in interdisciplinary learning communities, collaborative learning, and student self-assessment, and in ways assessment can be used to strengthen innovative efforts in undergraduate teaching and learning. Alongside her Washington Center involvement, Jean teaches part-time in Evergreen's Master's of Environmental Studies program.

As education specialist for the Malcolm Baldrige National Quality Award program, I have had the opportunity to work with many organizations looking at their evaluation and improvement. Today, I would like to tie together the perspective that Baldrige brings to institutional assessment with what you will be hearing at this conference.

The Malcolm Baldrige Approach and Assessment

by Sue Rohan

The Malcolm Baldrige National Quality Award was created in 1987 for businesses that are for-profit, hence clearly different in many ways from colleges and universities. However, lessons can be learned from what has happened with the program and its criteria, which have been widely accepted in this country and abroad. To identify and share best practices throughout the country, the Baldrige program uses a set of criteria. The criteria, a tool for self-assessment, are used by the program to identify best practices for performance excellence that can be shared for the benefit of other organizations.

Over the ten years of the Baldrige program, we have improved the criteria each year. We have learned about the value of self-assessment and the value of external comparisons. Today I will discuss the Baldrige approach to assessment that has been found to be effective and some of the results of the 1995 Education Pilot. I will tie these learnings to a strategy for listening to the presentations across this conference's Strand Three that will provide a context for the various approaches to assessment.

Why Self-Assess?

Why might an institution decide to undertake a comprehensive self-assessment? The time and resources expended can be considerable, especially if such a process is not already built into your institutional planning process. One reason for comprehensive self-assessment is meeting external requirements, particularly for public institutions. Governing boards sometimes request an evaluation of part or all of an institution; accrediting associations require periodic and systematic evaluations.

Another reason is that colleges and universities are facing increasing quality, cost, and marketplace challenges. Most

Once you have completed a self-assessment, know your strengths, and have prioritized needed improvements, you can plan an improvement strategy. You also have an opportunity to participate in an informed way through a common language (the Baldrige criteria) with other colleges, universities, and organizations in networks to arrange information exchanges.

leaders in both business and education believe that these challenges will intensify and become more complex due to such societal factors as changes in technology and the increasingly global economy. Assessment followed by corresponding improvement and innovation will help prepare an institution to respond to tomorrow's challenges.

Whether assessment is for the purpose of meeting external requirements or the result of an internal decision, that assessment can be a useful diagnostic tool to identify the strengths of the institution (those approaches on which you might wish to build) and the opportunities for improvement (those approaches not serving you as well as they could). The approaches could need enhancement, better alignment with other aspects of the organization, further deployment, or radical changes.

Once you have completed a self-assessment, know your strengths, and have prioritized needed improvements, you can plan an improvement strategy. You also have an opportunity to participate in an informed way through a common language (the Baldrige criteria) with other colleges, universities, and organizations in networks to arrange information exchanges. Such networking often provides insights regarding best practices that can help in your improvement initiatives.

The Concept of Excellence

As an institution begins a self-assessment, what is the "concept of excellence" underlying such an initiative? What is it that you are trying to achieve or that you are working toward? The concept of excellence built into the Baldrige criteria is that of demonstrated performance. Such performance has two manifestations: (1) year-to-year improvement in key measures and/or indicators of performance; and (2) demonstrated leadership in performance and performance improvement relative to comparable institutions and/or to appropriate benchmarks.

This concept of excellence has been selected because it places the major focus on teaching and learning strategies; it poses similar types of challenges for all colleges and universities regardless of their resources and/or the preparation/abilities of their incoming students; it is most likely to stimulate learning-related research and to offer a means to disseminate the results of such research; and it offers the potential to create an expanding body of knowledge of successful teaching/learning practices in the widest range of postsecondary institutions.

The focus on "value-added" contributions by the college/university does not presuppose a manufacturing-oriented, mechanistic, or additive model of student development. Nor does the use of a value-added concept imply that the institution's management system should include documented "procedures" or attempt to define "conformity" or "compliance." Rather, the performance concept in the Baldrige Education Pilot criteria means that the college or university should view itself as a key developmental influence (though not the only

one in a student's life) and that it should seek to understand and optimize its influencing factors, guided by an effective assessment strategy. For example, a university could improve its performance ratings by raising admission standards of the incoming class; but that is not the sort of performance improvement the criteria are seeking. The criteria focus more on what the college or university has done to add value to the learning and lives of the students given that the instructional process has been one of the key influences in the students' lives.

Key Characteristics of the Baldrige Education Pilot Criteria

Performance Results

The criteria are directed toward improved overall institutional performance results. The criteria focus principally on the key areas of college/university performance, given below. In the Baldrige program, performance results are a composite of the following:

- student performance
- student success/satisfaction
- stakeholder satisfaction
- institutional performance relative to comparable institutions
- effective and efficient use of resources.

Improvements in these result areas comprise overall college/university performance in the award program.

The use of a composite of indicators helps to ensure that strategies are balanced — that they do not trade off among important stakeholders or objectives. The composite of indicators also helps to ensure that institutional strategies bridge short-term and long-term goals.

Systems Approach

The Baldrige criteria support a systems approach to organization-wide goal alignment. The systems approach to goal alignment is embedded in the integrated structure of the award's criteria and the results-oriented, cause-effect linkages among the criteria items.

Alignment in the criteria is built around connecting and reinforcing measures, derived from the organization's strategy. These measures tie directly to the student/stakeholder value and to overall performance that relates to key internal and external requirements of the institution. The use of measures thus channels different activities in consistent directions without the need for detailed procedures or centralization of decision making or process management. Measures thus serve both as a communications tool and as a basis for deploying consistent overall performance requirements.

Such alignment, then, ensures consistency of purpose while at the same time supporting speed, innovation, and decentralized decision making.

Learning and Improvement Cycles

A systems approach to goal alignment, particularly when strategy and goals change over time, requires dynamic linkages among criteria categories and items that together foster systems learning. In the Baldrige criteria, action-oriented learning takes place via feedback between processes and results via learning cycles.

The learning cycles have four, clearly defined stages: planning, including design of processes, selection of measures, and deployment of requirements; execution of plans; assessment of progress, taking into account internal and external results; and revision of plans based upon assessment findings, learning, new inputs, and new requirements.

Assessment Strategy

Central and crucial to the success of the excellence concept in the Education Pilot criteria is a well-conceived and well-executed assessment strategy. The characteristics of such a strategy should include the following:

- Clear ties between what is assessed and the university's mission objectives. This means not only what students know but also what they're able to do.
- A strong focus on improvement — of student performance, faculty capabilities, and school program performance.
- Assessment as embedded, ongoing, with prompt feedback.
- Assessment, curriculum-based and criterion-referenced, that addresses key learning goals and overall performance requirements.
- Clear guidelines regarding how assessment results will be used and how they will *not* be used.
- Ongoing evaluation of the assessment system itself to improve the connection between assessment and student success. Success factors should be developed based on external requirements of graduates derived from the marketplace, other colleges and universities, and additional sources on an ongoing basis.

Education Criteria Purposes and Goals and Their Relation to the Conference Topics

The Education Pilot criteria are the basis for assessment and feedback to education organizations. The criteria have four additional purposes that could form a common foundation for the types of assessment you will hear about at this conference:

- to help improve institutional performance practices by making available an integrated, results-oriented set of key performance requirements;
- to facilitate communication and sharing of best practices information within and among institutions of all types based upon a common understanding of key performance requirements;
- to foster the development of partnerships involving educational institutions, businesses, human service agencies, and other organizations via related criteria; and
- to serve as a working tool for understanding and improving organizational performance, planning, training, and institutional assessment.

Education Criteria Goals

The criteria are designed to help colleges and universities improve their educational services through a focus on dual, results-oriented goals. These two goals are (1) provision of ever-improving educational value to students, contributing to their overall development and well-being, and (2) improvement of overall school effectiveness, use of resources, and capabilities.

These goals might also provide a basis for evaluating the ways in which various assessment and accreditation practices presented at this conference can be useful to your institution.

Criteria for Performance Excellence Framework

The education criteria are based on a set of core values and concepts, including learning-centered education, leadership, continuous improvement and organizational learning, valuing faculty and staff, partnership development, design quality and prevention, management by fact, long-range view of the future, public responsibility and citizenship, fast response, and

results orientation.

These core values and concepts are embodied in seven categories: Leadership, Strategic Planning, Student and Stakeholder Focus, Information and Analysis, Faculty and Staff Focus, Educational and Support Process Management, and School Performance Results.

The framework connecting and integrating these categories has three basic elements, as diagrammed in Figure 1 *(on the next page)*. Let's look at each of them, working from top to bottom.

Strategy and action plans are the set of student and other stakeholder-focused institutional-level requirements, derived from short- and long-term strategic planning, that must be done well for the organization's strategy to succeed. Strategy and action plans guide overall resource decisions and drive the alignment of measures for all work units.

System, the second part of the framework, comprises the six categories in the center of the figure that define the organization, its operations, and its results. Categories 1-3 represent the *leadership* triad; they are placed together to emphasize the importance of a leadership focus on strategy and students. Categories 5-7 represent the *results* triad; an institution's employees and its key processes accomplish the work of the organization that yields its results.

All institutional actions point toward a composite of performance results. The large arrow in the center of the framework links the leadership triad to the results triad, a linkage critical to college and university success. Furthermore, the arrow indicates the central relationship between leadership and school performance results. Leadership (category 1) must keep its eyes on the results (category 7) and must learn from them to drive improvement.

Information and analysis (category 4) is critical to the effective management of the institution and to a fact-based system for improving performance. Information and analysis serve as a foundation for the performance management system.

Criteria Structure

The seven criteria categories are subdivided into "items" and "areas to address." Each of eighteen items focuses on a major requirement. Items consist of one or more areas to address. Information for assessment is prepared in response to the specific requirements of these areas.

Let's look at these categories and the items associated with each:

1. Leadership. The leadership category addresses how senior leaders guide the university in setting directions and seeking future opportunities. It addresses how senior leaders create a leadership system that is based upon clear values and high performance expectations, and that addresses the needs of all stakeholders. The two items under leadership focus on:

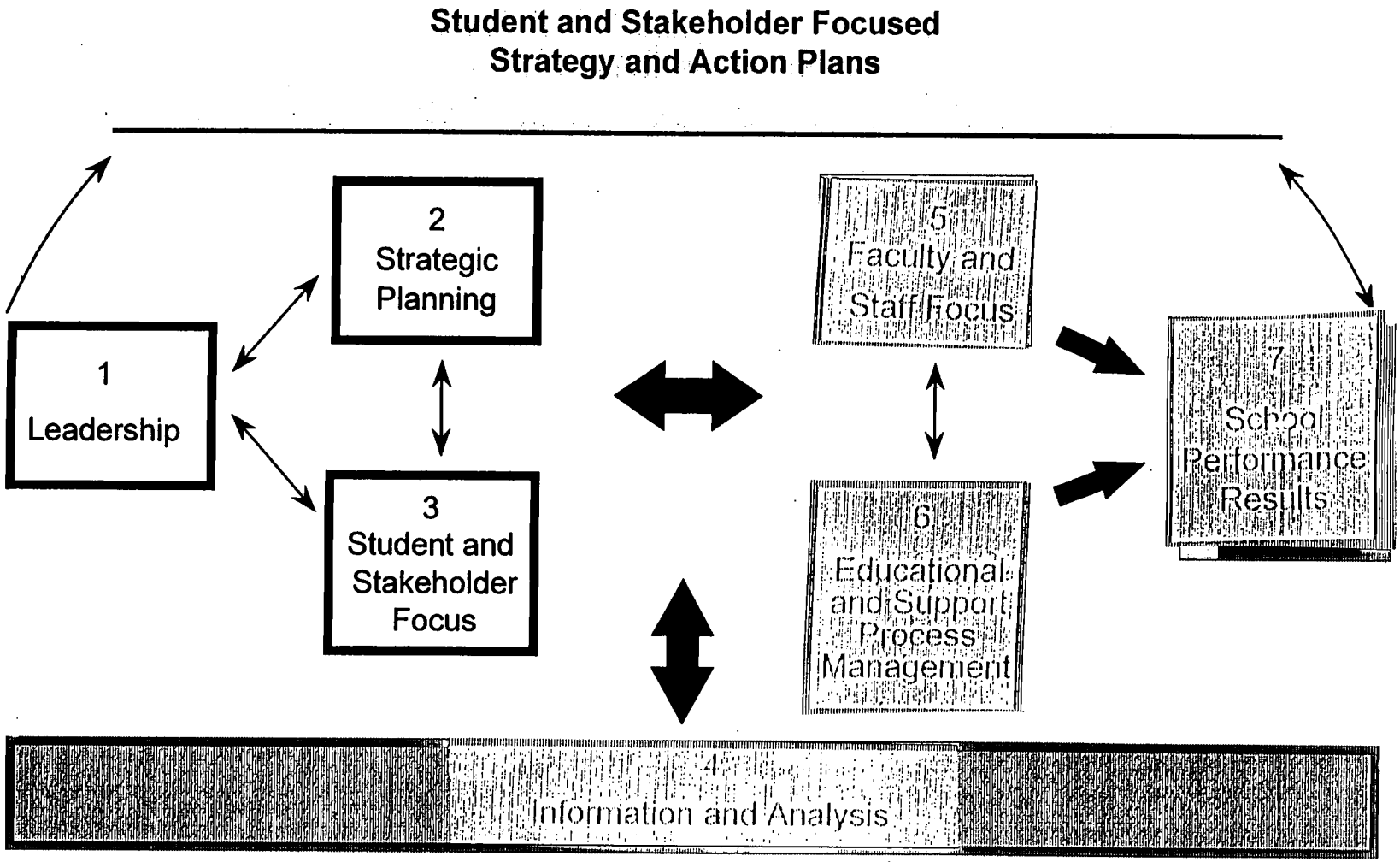
1.1 How senior leaders create values and expectations, set directions, project a strong customer focus, encourage innovation, develop and maintain an effective leadership system, effectively communicate this information, and effectively review and improve the system.

1.2 How the college/university integrates its values and expectations regarding its public responsibilities and citizenship into its performance management system, and how societal responsibility, including regulatory, legal, and ethical responsibilities and community involvement, are addressed.

We in higher education might learn from other organizations about successful

Baldrige Education Criteria Framework
A Systems Perspective

Figure 1



BEST COPY AVAILABLE

leadership systems. "Key excellence indicators" for leadership observed in manufacturing, service, and small business Baldrige Award recipients could well be modified and applied within higher education:

- strong customer focus
- high visibility
- set aggressive "leapfrog" goals
- leaders drive cycle time
- clear, easily remembered values
- managers as coaches
- a focus on continuous learning
- champion for company citizenship
- patient

2. Strategic Planning. This category addresses all aspects of organization-level planning and the deployment of plans. It includes primarily the development and deployment of key educational and other mission-related requirements, taking into account the needs of students and other key stakeholders. The strategic planning category examines how schools understand key student and stakeholder and societal requirements as input to setting directions; optimize the use of resources, ensure faculty and staff capability, and ensure bridging between short- and longer-term requirements; and ensure that plan deployment will be effective — that there are mechanisms to communicate requirements and achieve overall alignment.

The two items under strategic planning look at the strategy development process and school strategy:

2.1 How the institution develops its view of the future, sets directions, and translates these directions into a clear basis for communicating, deploying, and aligning critical requirements. Alignment refers to effective integration of faculty development, curriculum, instruction, and assessment.

2.2 How strategy and action plans are deployed. Also calls for a projection of the institutional performance. The main intent of the item is effective operationalizing of action plans, incorporating measures that permit clear communication and tracking of progress and performance.

Some key excellence indicators seen in Baldrige Award recipients include these:

- quality planning is business planning
- long-term horizon
- aggressive planning drivers (benchmarks) derived from study of world leaders
- covers products, services, processes
- key targets derived from customer requirements and market directions — current and future, deployed to all units
- links to suppliers and partners

3. Student and Stakeholder Focus. This criteria category explores how the higher education institution seeks to understand the needs of current and future students and other stakeholders on an ongoing basis. It stresses the importance of school relationships and of the use of an array of listening and learning strategies. Although many needs of stakeholders must be translated into educational services for students, the stakeholders themselves have needs that schools must also accommodate. A key challenge to schools is to balance differing needs and expectations of students and stakeholders and among stakeholders themselves.

3.1 How the institution determines the needs and expectations of its current and future students to maintain a climate conducive to learning for all students. Student needs should take into account information not only from

students but also from families, employers, and other schools, as appropriate. Student needs should be interpreted in a holistic sense to include knowledge, application of knowledge, problem solving, and learning skills.

3.2 How the college/university determines and enhances the satisfaction of its students and stakeholders to build relationships to improve educational services and to support related planning. The item calls for information on how the organization provides for effective relationships with key stakeholders to enhance its ability to improve educational services. It also addresses how the school determines student and stakeholder satisfaction and dissatisfaction for use in improving the school's ability to improve educational and support services. A critical part of this process is how the school's measurements capture key information that bears upon students' motivation and active learning.

Key excellence indicators observed in Baldrige Award recipients include these:

- market knowledge
- proactive customer systems
- use of all listening posts, such as surveys, product/service follow-up, complaints, customer turnover, and all staff
- knowledge of requirements of market segments
- surveys go beyond current customers
- front-line empowerment
- strategic infrastructure support for front-line employees
- focus on relationship management and enhancement
- attention to hiring, training, attitude, and morale of all employees
- high levels of satisfaction, customer awards

4. Information and Analysis. Information and analysis is the main point within the criteria for all key information to effectively manage the organization and to drive performance improvement. It addresses all basic performance-related information and comparative information, as well as how such information is analyzed and used to optimize school performance.

4.1 Selection, management, and use of information and data to support overall organizational goals, with strong emphasis on action plans and performance improvement. Key factors in the effective selection and use of data include (1) the main types of information and data and how each type relates to key school processes and action plans; (2) how information and data are made available to all users to support effective day-to-day management and evaluation of key processes; (3) how key user requirements — rapid access, reliability, and confidentiality are met; and (4) how all aspects of data and information — selection, deployment and user requirements — are evaluated, improved, and kept current with changing needs.

4.2 External drivers of improvement — data and information related to best practices, new practices, and to performance of comparable higher education institutions and other organizations. The major premises underlying this item are (1) colleges and universities need to "know where they stand" relative to comparable schools and/or other organizations; (2) comparative and benchmarking information often provide impetus for major change and improvement, and might signal changes taking place in educational practices; and (3) organizations need to

understand their own processes and the processes of others before they compare performance levels.

4.3 Organization-level analysis of overall performance — the principal basis for guiding processes toward key results. Five key aspects of school performance are addressed: (1) student and student groups; (2) school programs; (3) student, student group, and school programs relative to comparable schools; (4) school operational performance; and (5) school operational performance relative to comparable schools.

Analyses that schools carry out to gain understanding of performance vary widely. Selection depends upon many factors, including type of educational institution, size, and relationship to other organizations.

Examples of such analyses include trends in key indicators of student motivation such as absenteeism, dropout rates, and use of educational facilities; test performance trends for students, segmented by student groups; relationships between in-school outcomes and performance and longer-range outcomes — in other schools or in the workplace, for example; activity-level cost trends in school operations; student utilization of learning technologies and/or facilities versus assessment performance; relationships between student background variables and outcomes; relationships between student allocation of time to activities and projects and academic performance; and percentage of students attaining industry-based and/or profession-based skill certification.

Overall, item 4.3 represents the basis for judging institutional effectiveness, including use of all resources.

Key excellence indicators for information and analysis include these:

- quantitative orientation
- focus on actionable data
- multiple measures
- inter-linking measures — internal and external
- wide deployment and accessibility
- strong analysis capability
- benchmark best-in-class, within and outside of industry

5. Faculty and Staff Focus. Category 5 addresses all key human resource issues and practices in an integrated way, aligned with the school's mission and strategy. Three items in this area are:

5.1 How work and job design, compensation, and recognition approaches enable and encourage all faculty and staff to contribute fully and effectively.

5.2 How the school develops faculty and staff via education, training, and other developmental approaches, formal and informal.

5.3 Work environment and work climate that support and enhance the well-being, satisfaction, and motivation of faculty and staff.

Key excellence indicators for the Faculty and Staff Focus criteria observed in Baldrige Award recipients include these:

- integration with overall business planning
- "internal customer" focus
- comprehensive training and education
- individual and organizational learning linked
- empowerment, cross-training
- team and individual recognition
- lower turnover, accidents, absenteeism
- commitment to employee satisfaction, motivation, and well-being

6. Educational and Support Process Management. This criteria category addresses all key school processes. It considers requirements for efficient and effective process management, including effective design, evaluation, continuous improvement, and a focus on high performance.

6.1 How the organization designs, introduces, delivers, and improves its educational programs and offerings. This item also examines organizational learning, through a focus on how learnings in one school work unit are replicated and added to the knowledge base for other school units.

Four aspects of education design are included: (1) how student educational and well-being needs are addressed, with a strong focus on active learning and taking into account varying learning rates and styles; (2) how sequencing and offering linkages are addressed; (3) how design includes a measurement plan that makes use of formative and summative assessments; and (4) how the school ensures that faculty are properly prepared.

Design approaches might differ appreciably depending upon many factors including school mission, as well as student age, experience, and capability. Formative and summative assessments need to be tailored to the offering and program goals, and might range from purely individualized to group-based.

This item also calls for information on program and offering delivery. Offering delivery refers to all strategies used to engage students in learning. Examined are the observations, measures, and/or indicators used and how these are used to provide timely information to help students and faculty.

6.2 How the organization designs, maintains, and improves its support processes. Support processes are those that support the school's overall education activities and operations. This includes learner support services such as counseling, advising, placement, tutorial, and libraries and information technology. It also includes, as appropriate, recruitment, enrollment, registration, accounting, plant and facilities management, secretarial and other administrative services, security, marketing, information services, public relations, food services, health services, transportation, housing, bookstores, and purchasing.

Key excellence indicators observed in Process Management among Baldrige Award recipients include these:

- products, services, and business processes
- quality in design — products, services, processes
- focus on cycle time and productivity
- integration of prevention, correction, and improvement with daily operations
- supplier partnering

7. School Performance Results. The seventh category provides a results focus for all school improvement activities, using a set of measures that reflect overall mission-related success. Data called for are the major ingredients in earlier item 4.3, which is intended to identify causal connections to support improvement activities, planning, and change. Overall, the four items in this category should provide a comprehensive and balanced view of the school's effectiveness in improving its performance, now and in the future.

7.1 Principal student performance results

based upon mission-related factors and assessment methods. Critical to understanding the purposes of this item are that (1) student performance should reflect holistic and mission-related results; (2) current levels and trends should be reported — the former to allow comparisons with other schools and/or student populations, and the latter to demonstrate year-to-year improvement; and (3) data should be segmented by student group(s) to permit trends and comparisons that demonstrate the school's sensitivity to education improvement for all students.

Overall, this item is the most important one, as it depends upon demonstrating improvement by the school over time and higher achievement levels relative to comparable schools and/or student populations.

Item 7.1 depends upon appropriate normalization of data to compensate for initial differences in student populations. Although better admission criteria might contribute to improved education for all students, improved student performance based entirely upon changing students' entry-level qualifications does not address its requirements.

7.2 Trends and levels in student and stakeholder satisfaction based on relevant measures and/or indicators, and these results compared with comparable schools. Effectively used, satisfaction results provide important indicators of school effectiveness and improvement. Effective use entails understanding the key dimensions of satisfaction and dissatisfaction, recognition that satisfaction and dissatisfaction with school services and/or performance might differ among student and stakeholder groups, and that their level of satisfaction/dissatisfaction might change over time, based on longer-term perspectives.

7.3 Human resource results — those related to well-being, development, satisfaction, and performance of faculty and staff. Results reported could include safety, absenteeism, turnover, and satisfaction. School-specific factors might include those created by the school to measure progress against key goals. This item calls for comparative information so that results can be evaluated relative to comparable institutions.

7.4 Key performance results not covered in items 7.1-7.3 that contribute significantly to the school's mission and goals. This item encourages the use of any common or unique measures the school uses to track performance in areas of importance to the school's mission and goals.

Appropriate for inclusion are measures of productivity and operational effectiveness, including timeliness; results of compliance and improvement in areas of regulation, athletic programs, etc.; improvements in admission standards; improvements in school safety and hiring equity; effectiveness of research and services; school innovations; utilization of school facilities by community organizations; contributions to community betterment; improved performance of administrative and other school support functions; cost containment; and redirection of resources to education from other areas.

The item calls for comparative information so that results reported can be evaluated against other organiza-

tions. Such data might include results of surveys and peer ratings.

Key excellence indicators for School Performance Results include these:

- broad array of customer satisfaction measures, including segmentation
- broad base of improvement trends and/or excellent performance, including products, services, internal operations, cycle time, and productivity
- results for employees (the "internal customer") emphasized
- results "benchmarked" to leaders
- results of financial and marketplace performance tied to improvements
- improvements in supplier performance

The Evaluation System

When Baldrige Award examiners evaluate a written application, they provide feedback on an organization's strengths and opportunities for improvement along three evaluation dimensions: approach, deployment, and results.

Let's review these dimensions and then look at how educational institutions fared in the 1995 Baldrige pilot program with education and health care organizations.

Approach

The approach dimension refers to how the item requirements are addressed — the method(s) used to meet mission-specific requirements. The factors used to evaluate approaches include these:

- Appropriateness of the methods to the requirements.
- Effectiveness of use of the methods, including degree to which the approach is systematic, integrated, and consistently applied; embodies evaluation/improvement/learning cycles; and is based on reliable information and data.
- Evidence of innovation and/or significant and effective adaptations of approaches used in other types of applications or sectors.

Deployment

Deployment refers to the extent to which the approach is applied to all appropriate parts of the organization. The factors used to evaluate deployment include these:

- Use of the approach in addressing organizational requirements and Baldrige criteria requirements for each item.
- Use of the approach by all appropriate work units.

Results

Results refers to outcomes in achieving the purposes given in the criteria item. The factors used to evaluate results include these:

- Current performance.
- Performance relative to appropriate comparisons and/or benchmarks.
- Rate, breadth, and importance of performance improvements.
- Demonstration of sustained improvement and/or sustained high-level performance.
- Linkage of results measures to key performance measures identified by the organization.

Results of the 1995 Education Pilot

Let's look at the evaluation results from the 1995 Education Pilot. Nineteen institutions participated in the program, about half elementary/secondary schools and the other half postsecondary institutions.

About two-thirds of the postsecondary applicants were full universities. No technical or community colleges chose to partici-

pate. Of all the applicants, only two universities were private; all others were public and all were not-for-profit. The number of employees in the institutions gives a feel for their size. The institutions ranged from 75 to 1,200 employees, with an average of 455. The number of sites ranged from one to five.

Applicants for the Baldrige Award can receive up to 1,000 points. About 90% of the Education Pilot applicants scored between 0 and 450 points. In comparison, the majority of the Business Award applicants scored in the 451- to 650-point range.

Typically, applicants with scores below 450 have the beginnings of systematic approaches to performance quality. Such organizations have not yet fully identified all their customers and their key requirements. They still react to problems, rather than having a general improvement orientation. They have major gaps in deployment. Their attention to organizational performance and improvement over and throughout the entire organization is not consistent; some department programs are well developed, while others have not even started.

Organizations with scores of less than 450 typically are in the early stages of developing trends for their results measures. Good performance is displayed in only a few areas. Trend data consist of only a few data points over a year or two, or data with no consistent pattern to the improvements. Finally, the results for many areas of importance to key requirements are not reported.

Figure 2 *(on the next page)* displays the average scores in each of the seven Baldrige categories for applicants to the Business Awards in manufacturing and in service and for Education and Health Care Pilot applicants. The order of the categories shown here is for the 1995 criteria, which as you may know was changed in the 1997 and 1998 criteria.

You can see similarity to the patterns of the lines, particularly in Education, Health Care, and Service, showing that education and health care organizations perform similarly and are more like businesses in the service sector than in manufacturing.

For education, health, and service, category 1, Leadership, and category 4, Human Resources, have higher average scores, while category 2, Information and Analysis, and category 6, Performance Results, have more room for improvement.

Figure 2 also demonstrates the difference in the maturity of the business organizations and of education and health care organizations. The applicants in the Education and Health Care Pilot on average are in the 20%-40% range, indicating beginnings of systematic approaches, gaps in deployment, and erratic results measurement. In contrast, the service organizations have average category scores in the 40%-60% range, representing systematic processes, none or few gaps in deployment, and many trend results related to the applicants' key requirements. These differences between service and pilot applicants are not unexpected. We anticipate over time that education and health care scores will rise as such organizations mature toward performance excellence.

What We Learned, and What It Might Mean for This Conference

We can build on what we learned from the 1995 Education Pilot to develop some strategies to maximize self-assessment and accreditation processes.

- It is helpful for the leadership, faculty, and staff to understand and agree upon

1995 Average Category Scores

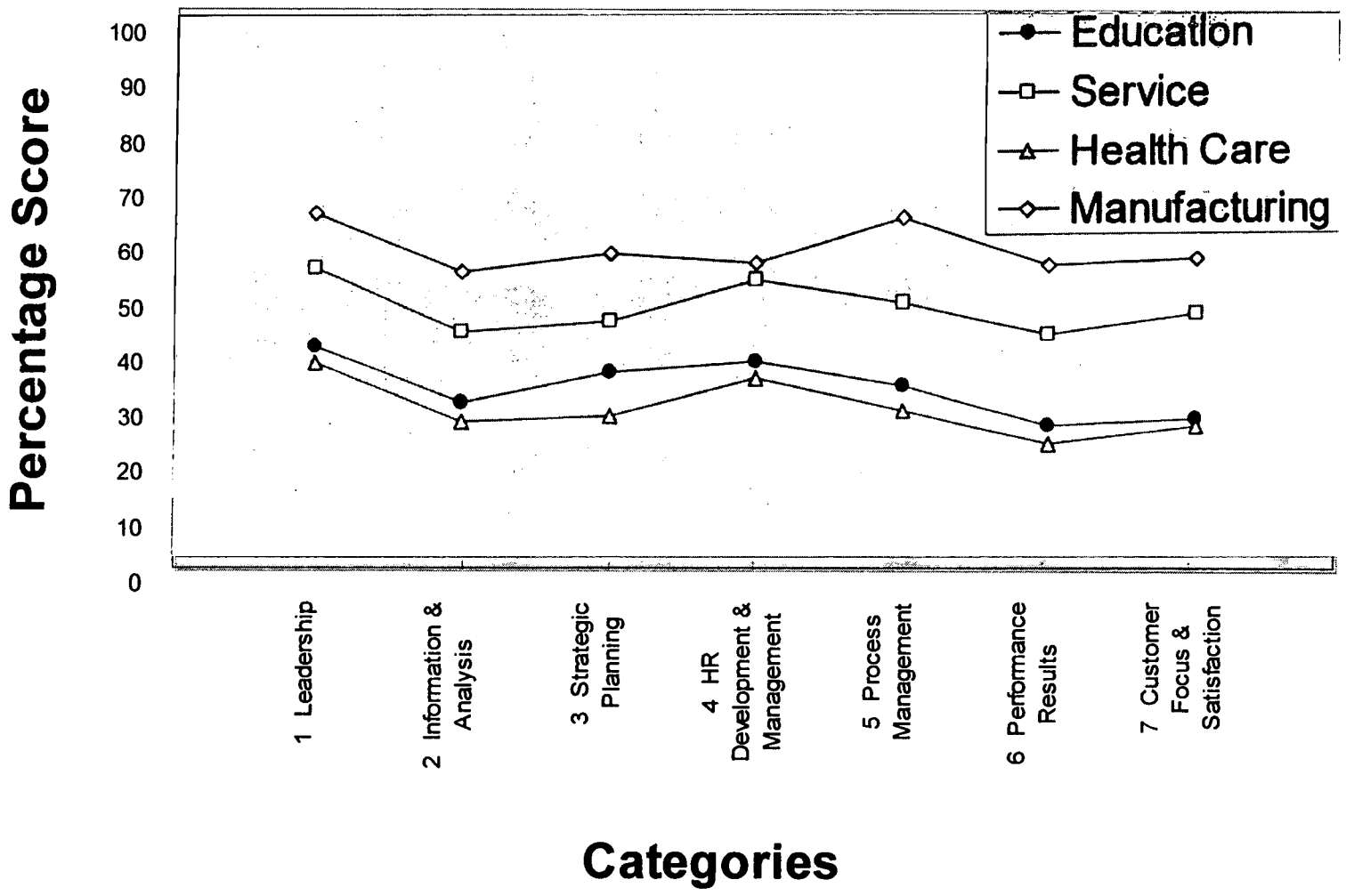


Figure 2

BEST COPY AVAILABLE

the institution's mission, to determine what the institution is trying to accomplish before starting an assessment.

- Having in place an organization-wide strategy with a set of action plans, specific goals, measures, and resource needs provides a basis for the assessment of approaches and deployment.
- Measuring the outcomes in a composite of areas identified as strategic and related to the specified goals provides a basis for evaluating results.
- Results need to include current and past performance relative to appropriate comparisons and benchmarks.
- Using a basic self-assessment, an organization can evaluate progress toward goals and improve plans, processes, measures, and results, thus making the assessment process a useful and ongoing organizational improvement tool.
- Assessment should help determine where an organization was, where it is now, and where it is going.
- An organization's accomplishments can best be evaluated in the context of a comparison with competitors, peer institutions, and benchmarks within and outside of education.

A Fresh Perspective

As you evaluate various assessment approaches over the next few days, it may be helpful to keep in mind the integrated framework of the Baldrige approach and its seven categories of Leadership, Strategic Planning, Student/Stakeholder Focus, Information and Analysis, Faculty and Staff Focus, Educational and Support Process Management, and the composite of Performance Results.

The Baldrige approach views the organization as a total system, and to evaluate one component without the rest might not portray an accurate picture or provide actionable feedback. It is important to link strategies, methods, approaches, deployment, and results to learn about the cause-and-effect relationships in an effort to improve.

Is self-assessment worth it? Can accreditation be a useful process?

Using a well-developed assessment approach, Baldrige applicants experience improved communication throughout the organization, better alignment of resources, and progress toward excellence.

You will be hearing many excellent presentations on assessment and accreditation during this conference. Strand Three sessions address various parts of what is included in the Baldrige framework — for example, use of satisfaction measures, measuring performance outcomes, using expectations of external stakeholders, improvement in student affairs, benchmarking options, and administrative and student support options. Baldrige may provide a foundation for bringing together many different ideas that will be presented here.

I hope that you find the Malcolm Baldrige National Quality Award framework useful as you explore the many facets of self-assessment and accreditation. ●

.....

Sue Rohan began work on a federal education award program in 1994, as the education specialist for the Malcolm Baldrige National Quality Award Office at the National Institute of Standards and Technology. Before she joined the Baldrige program, Rohan was a senior consultant for quality improvement at the University of Wisconsin System, where she

worked with its twenty-six campuses to improve quality. In Wisconsin, she rewrote the Baldrige Business Criteria for use by the University of Wisconsin, which contributed to development of the Malcolm Baldrige National Quality Award Education Criteria.

Rohan was a member of the Wisconsin State Legislature from 1985 to 1992, where she worked to promote quality management practices in state government. Previously, she had been a teacher of learning-disabled students, an educational diagnostician, and an elected representative of the Madison, Wisconsin, teachers union. Her blend of experience gives her a unique perspective on issues of educational evaluation, quality, and stakeholder satisfaction.

Assessment of Programs and Units

by Jon F. Wergin

When I was first asked to take on the role of "strand introducer," I thought it sounded a little sinister — as if I were being asked to infect the conference with a foreign agent of some sort. But after I thought about it some more, I realized that the metaphor probably wasn't far off the mark — that part of my task should be to be a bit provocative, to get under the skin of people. And so I'll try to do that today. I'll begin with some commentary on the state of program review and assessment, offer some opinions on what accounts for the state it's in, and then end with some challenges we need to face in order for program review to fulfill what I think has so far been a largely wasted potential.

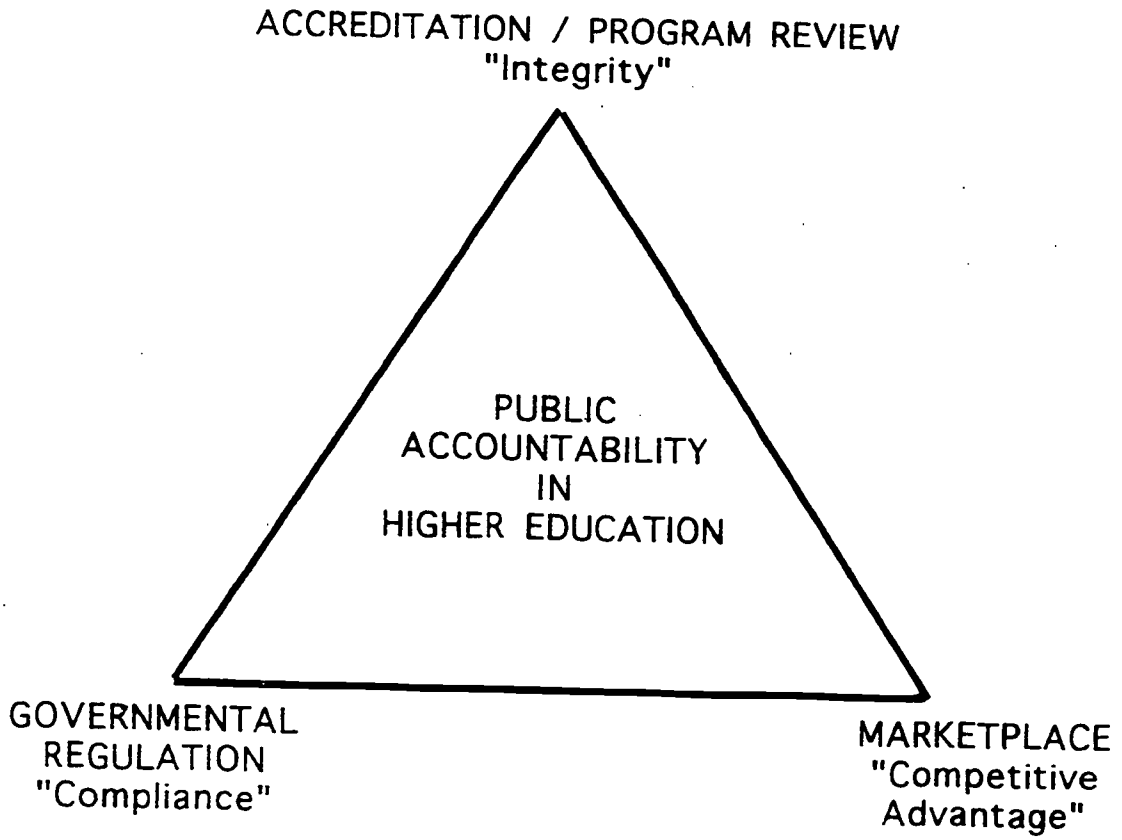
Let me say a bit about my own perspective on all of this. I've played multiple roles throughout my academic career: professional staff, consultant, program administrator, and most recently and currently member of the teaching faculty. I've looked at the evaluation of programs from all sides of the table. In the past few years I've become especially interested in how academic departments work (in some quarters, I know, that's an oxymoron) and, in particular, how departmental cultures affect issues of evaluation, both of individual faculty members and of the department as a whole. I've become convinced that departmental cultures are the key to effective program assessment. Unfortunately, more often than not, departmental cultures are substantial *barriers* to effective program assessment.

Think about these two terms for a moment: "program review" and "accreditation." In most departments these are topics that make the eyes of faculty roll to the back of the head. As Ted Marchese observed several years ago, program review is a process much of the academic world could imagine doing without. Why is this? Why have these activities become so ritualistic in most places? Especially since the hallmark of both institutional program review and accreditation is a self-study, which by definition calls upon the very qualities of analysis and reflection that academics value most highly?

Consider this simple diagram *(shown on the next page)*.

Higher education maintains its public accountability and assures its usefulness to society in three ways:

Figure


ACCREDITATION / PROGRAM REVIEW
"Integrity"
PUBLIC
ACCOUNTABILITY
IN
HIGHER EDUCATION
GOVERNMENTAL
REGULATION
"Compliance"
MARKETPLACE
"Competitive
Advantage"

- One is governmental regulation (which includes not only federal and state government but also state coordinating and governing boards). This corner of the triangle exists to ensure that higher education institutions are fiscally and socially responsible, that they meet appropriate safety and health standards, and that they offer educational programs that aren't unnecessarily duplicative. The goal of regulation is *compliance*.
- Another force for public accountability is the marketplace. Particularly with the advent of technology and distance learning, the competition for students among educational providers is increasing. Institutions that fail to adjust to a changing market put their own health and survival at risk. The goal of the marketplace corner of the triangle is *competitive advantage*.
- At the top of the triangle is program review. I put it at the top for a reason: Of the three forms of public accountability, this is the only one that focuses on the quality and *integrity* of the work itself, and it's the only one over which the institution and its faculty have any direct control. The collective faculty have traditionally been the ones responsible for maintaining program quality, and no one wants to leave that function to the government or the marketplace.

So here's the paradox: The form of public accountability in which the institution has — or should have — the greatest vested interest is also usually the weakest. So, again, why is this the case?

First of all, I'm afraid that we've succumbed to a compliance mentality in higher education. The questions driving many program reviews are "theirs," not "ours." The review is on someone else's agenda: higher administration, governing board, professional or disciplinary society. Most faculty accept the necessity of program review, but don't generally see it as a process that will affect their own professional practice, at least not in a positive way.

A second and related problem is that most program reviews are one-shot affairs, not well integrated into the life of the institution. Unless the program review has been triggered by an administrative action that threatens the program's status quo, the process often unfolds in a way that allows the participants to get through the process with a minimum of aggravation. The self-study is given over to selected staff and a few faculty members who, if they are lucky, will be given some release time to conduct the study and write the report. The whole process becomes tedious, time-consuming, and too often ultimately of little or no consequence.

Because the focus is backward (on what has already happened) rather than forward (on what is possible), the review is a ritual. The opportunity for critical reflection — a chance to put our strong academic values of systematic inquiry and questioning of assumptions to use — is lost in the desire to get the thing done.

Those of you who are active in the assessment movement have probably heard these points before. For the last ten years at least, one of the chronic issues has been the problem of making assessment meaningful and useful to the faculty in the trenches. But I'd like to offer a third reason for the widespread perception that program review is of little consequence to the life of the institution, and it goes back to a point I made earlier about departmental culture. The point is this: The faculty culture in most departments is individualistic and highly privatized. One anonymous pundit has said

Because the focus is backward (on what has already happened) rather than forward (on what is possible), the review is a ritual. The opportunity for critical reflection — a chance to put our strong academic values of systematic inquiry and questioning of assumptions to use — is lost in the desire to get the thing done.

this: "Academic departments are clans of arrogant experts seeking to sustain individual privilege at the expense of institutional goals." An overstatement? Sure. But I'll bet that it isn't much of a stretch to identify departments like these at your own institution. The problem is that evaluation devolves to the individual, not to the unit. Faculty are rewarded on the basis of their contributions to their profession or discipline, not to their institutions.

Furthermore, as Jim Fairweather, of Michigan State, has pointed out, when units *are* evaluated, they are normally judged on the basis of the sum of the performances of individual faculty — scholarly productivity, for example — not by measures of the unit's contribution to a larger good. The emphasis continues to be on individual merit, not on collective worth to the mission of the institution. As a consequence, there's little faculty investment in activities that require collective action, and the consequence of *that* is captured vividly in this private communication from an official of one of the regional accrediting associations:

> *The place of faculty [in program review] is uncertain. They carry little credibility with presidents, and seem increasingly unprepared to carry out responsibilities in shared governance. They don't seem to be creative players in preparing higher education for the future. Sometimes I wonder: Are faculty willing, let alone equipped, to share in the current transformation of higher education?*

Yikes.

This is a shame, if true. Program review and accreditation are the most public way for higher education to maintain a core set of values that has served us well: those values include autonomy, self-governance, and the pursuit of knowledge in a way that is unfettered by questions of efficiency or popularity. Program review remains the *only* peer-based mechanism for evaluating quality. We abrogate our responsibilities for peer review at our peril. If we faculty ignore this corner of the quality triangle, we risk having governmental and marketplace forces take over (as, some might argue, they are already doing).

Four Challenges

Given this uncomfortable scenario, how might program review become more useful? What are the challenges? I'll suggest four and then invite you to add a few of your own. For each I'll first pose the challenge and then suggest a central question you might ask as you wend your way through the ideas presented in this program strand.

Defining Quality. The first challenge is to get clear about what "quality" means. I feel a little embarrassed about saying this, to tell you the truth. This is an issue that should have been settled long ago, and maybe it has. Maybe I'm the only one who's still confused. But it seems to me that too many conversations about assessment proceed from the assumption that we have shared definitions of quality, and I just don't think that's true.

Here are two definitions of quality that have long outlived their usefulness: One is

the "transcendent" view, which holds that because the academy is the keeper of wisdom, quality is whatever we define it to be. Society's interests are necessarily served by advancing our own. This point of view is increasingly seen as socially irresponsible, and deservedly so.

The problem, in my opinion, is that the transcendent view of quality has been widely replaced by a second, "marketplace" view, which holds that quality is whatever we do that makes our customers (read *students*) happy. But too much responsiveness may itself be socially irresponsible. As the sociologist Everett Hughes once observed about the accountability of physicians: "A doctor who is too responsive to his patients is called a quack." As Larry Braskamp has observed, there's a difference between being respon*sive* and respon*sible*. Being responsible means working for the common good, which includes both addressing the expressed needs and priorities of those we serve *and* upholding the principles of academic freedom, the "free search for truth and its free exposition" (AAUP 1940).

Defining quality is thus a matter of negotiating interests. The criteria used to define a "quality program" are multidimensional, and will vary according to who the stakeholders are.

Faculty and administrators tend to mention different sets of criteria. Faculty members focus on such qualities as faculty credentials, fiscal resources and facilities, size of the faculty and student body, and degree of student involvement and quality of effort. Administrators focus on such things as enrollment demand, program centrality, and employability of graduates. Both groups, thankfully, will usually mention student learning. The point is that different stakeholder groups have different notions of what quality is; the notions are overlapping, to be sure, but they're not the same.

Furthermore, characteristics that define a "quality" program in one institution are not the same as those that define quality in another institution having a different educational mission. I'm not trying to suggest here that defining quality is a hopeless proposition — only that in order for program review to work, diverse interests must be recognized and negotiated. Note that I use the term "negotiated," not "catered to." A negotiated view of quality means that we need to recognize the constructive tensions between scholarship and social relevance, between faculty independence and collaboration with the larger community, and between the roles of social critic and social ally. What these tensions all boil down to is that faculty tend to think of quality in terms of excellence, or intrinsic merit, while external stakeholders tend to think of quality in terms of fitness for use, or worth.

Thus the $64,000 question, one that I'd like to see more guidelines for program review ask, is this: How well is the program pursuing excellence while at the same time delivering value?

Asking the Questions. A second challenge, which follows from the first, is how to make program review useful for answering multiple stakeholders' questions, including most specifically those of the faculty. It's hard to get faculty engaged in the process, or critically reflective about the results, when they're strictly answering someone else's questions, particularly when the data they're collecting involves counting things. One of the reasons why the Harvard Assessment Seminars have been so successful is that their work was organized around questions faculty found useful and intellectually interesting. For example, one of the most

effective conversation-stoppers in departmental faculty meetings is one that asks, "How can we improve our undergraduate teaching?" Compare that with a different question, one that asks, "What might we do to better prepare our students for a rapidly changing world?"

And so the second question I'd like to see asked more often is, "To what extent does the process encourage interaction around intellectually meaningful topics?" In other words, how well does program assessment better inform discussion about things people care about?

Making Peer Review Real. A third challenge is how to make program review real *peer* review. Peer review is a potentially powerful and positive force, but only if we think about the nature of assessment in a particular way. The Latin root of the word *assessment* is *assidere,* which means to "sit beside." When you think of assessment in this way, what comes to mind? To me it implies dialogue and discourse — understanding the other's perspective before making any judgments. Unfortunately, the more common image is of assessment as "standing over," which implies something very different indeed. Good peer review is a two-way conversation, one that challenges one's own perspectives and assumptions by understanding the perspectives and assumptions of others.

And so the third question I'd invite you to ask of a program review process is, "Do the results of program review resemble a conversation — or a briefing?"

Changing Cultures. The fourth and last challenge I want to suggest to you is hardest of all: How to shift the culture of academic programs from *individual* to *collective* accountability. Let me begin with a disclaimer: I've been studying collective responsibility in higher education for about five years now, and I'm afraid I might have fallen victim to a common fallacy — that what I'm studying contains most of the solution to what is wrong with higher education! But consider this: Over and over again, those of us who study academic departments and programs find a central variable that distinguishes the truly successful ones, and that variable is the degree to which departmental faculty take collective responsibility for the quality of the work they do. In my own work I've studied academic departments in about three dozen institutions, and I've come to roughly the same conclusion. Frankly, I don't know how program assessment could possibly be effective without a sense of collective responsibility, a sense that "this is our program, and we're all responsible for it."

But I've also found that "collective responsibility" as a concept is a lot easier to embrace in principle than in practice. Typically, departments that move in this direction do so in four stages, each successively more difficult than the one preceding it. First, they focus the mission of the program, usually as a response to an external threat of some sort. The challenge here is to maintain a mission focus after the immediate threat passes and the temptation is to return to business as usual. The second step is to get faculty to work together, and here the challenge is to develop a sense of interdependence and mutual accountability. The third stage is to develop differential faculty roles, in which work of the department is negotiated in ways that optimize the individual interests and talents of its members. The challenge of this step is to work through the very real fear that faculty *roles* will change, but faculty *rewards* will not. The fourth and most difficult step is to decide on new rules for evaluation of the unit as a whole. Shifting the focus of evalu-

ation from the individual to the unit requires a huge cultural change, marked by a willingness by both faculty and administrators to openly negotiate criteria and standards by which that unit will be judged. Few departments or academic administrators have the stomach for this; most seem to prefer, by default, the "black box" approach to evaluation, in which decisions about resource allocation are made ad hoc, behind closed doors, using criteria known only to the deal makers.

Program assessment has enormous potential as a way to open up this process and create the sort of cultural change I've described, but so far that potential is largely unrealized. And so the fourth question I would invite you to pose is, "Does program assessment support and encourage the development of cultures of collective responsibility?"

A Conference Plan

I invite you to reflect upon the challenges identified here this afternoon as you work your way through the sessions in this program strand. As you encounter success stories in program assessment and review — and there are many at this conference — consider how those campuses have dealt with the issues we've raised. See if you can discern any *other* challenges embedded in these cases.

And finally, consider how their successes might be transferable to your own institution. ●

.....

Jon Wergin is professor of educational studies at Virginia Commonwealth University. He teaches courses in adult and higher education, and coordinates the Preparing Future Faculty program for the VCU School of Graduate Studies. He has won awards from the School of Education for both scholarship and teaching.

Wergin is the author or coauthor of three recent books and numerous articles on such topics as the education of professionals, program evaluation in higher education, and the academic department as an agent of change. He is former divisional vice president of the American Educational Research Association; he has served as external evaluator of two national centers for research in higher education; and he has been a consultant in evaluation and organizational change to more than three dozen colleges, consortia, and professional associations. In 1992, he joined AAHE as interim director of its Forum on Faculty Roles & Rewards. He is currently a senior scholar with AAHE's New Pathways II project, and a consultant for the Pew Charitable Trusts on a survey of quality-assurance practices in academic departments.

Accreditation and Quality Assurance: Ambivalence and Confusion

by Judith S. Eaton

This morning I want to describe the Council for Higher Education Accreditation (CHEA), talk about faculty and accreditation, explore some dimensions of accreditation and quality assurance that affect us all, and offer some ideas for you about changing and about staying the same.

CHEA, an organization formed just two years ago, provides national coordination of voluntary accreditation. More than 3,000 colleges and universities are institutional members of CHEA, and approximately sixty-five accrediting organizations and higher education associations are organizational members. CHEA had an excellent second year for a young organization, whether measured by our growing revenues and membership, our progress on quality-assurance issues in higher education reauthorization, or the establishment of a research and policy capacity that provides a foundation for CHEA's national voice.

CHEA's intent is to provide leadership in ideas about quality assurance, advocacy for voluntary accreditation, and service to institutions and accreditors. As with any new and self-reflective organization, we have worked hard to address our organizational values, to identify what is important to us. CHEA is particularly concerned with reform and innovation in quality assurance, keeping student achievement always before us as the principal reason for being, and placing emphasis on the results of our higher education efforts.

The Faculty and Accreditation

Even after a short ten months on the job, I am convinced that we need more faculty investment and involvement in accreditation. We need more faculty presence in accreditation review, developing innovative reviews and, especially, paying more attention to student learning and achievement.

Accreditation is too important to be dismissed by faculty (as it is by some) as an "administrative activity" not worthy of faculty attention. It is too important to be another arena in which anti-administrator and anti-faculty sentiment play out. I have no patience with administrators who claim

that faculty "do nothing," just as I lack patience with faculty who claim that administrators "do only the wrong thing." We are all — faculty and administrators alike — challenged to support the values of higher education and provide leadership for change. The important work of accreditation takes all of us.

What if we woke up one morning to no accreditation? Would it make a difference to us? I submit that, yes, it would. I would guarantee you the presence of more federal and state control of higher education — and you would not like such control. I would guarantee you renewed and expanded assaults on institutional and faculty autonomy. Without accreditation, who would speak for our values and beliefs? There are other voices, to be sure, but the voice of self-regulation adds significant substance and weight.

The faculty role in accreditation must be grounded in the critical issues facing accreditation and must be viewed in the context of the responsibility of all educators for self-regulation. I believe that we limit ourselves — all of us — if we consider this role in isolation.

I now turn to a consideration of some dimensions of accreditation and quality assurance, issues of concern to faculty as well as others involved in accreditation.

Accreditation and Quality Assurance: Ambivalence and Confusion

At CHEA, I am struck by two features of many discussions about accreditation and quality assurance: the ambivalence about accreditation and the confusion about quality. The ambivalence about accreditation is plain to see: Almost everybody criticizes accreditation, but almost no one wants to do away with it. It appears to have value to many as a certificate of membership in the academic club of institutions — a rather special place — in spite of the criticism. The confusion about quality is also easily discernible: Everyone is for quality. Who would be against it? But there is little agreement on what *quality* means and how to use the concept.

Ambivalence About Accreditation: Why Do We React This Way?

First, we don't like — and we do like — accreditation. It is easy to criticize it — big, elastic, inefficient, and relying heavily on volunteer activity. And, the accreditation constituency is comparatively small, conserving in approach, and using language that, intentionally or unintentionally, is less than precise. We are not always clear, for example, about what is meant by "improvement" and "institutional integrity." Accreditation has also absorbed some government oversight function, and we don't like this examination.

On the other hand, the principles of voluntary self-regulation are valued. Affirmation of value is itself valued. We prefer to establish this valuing independent of the government and the market, and we believe voluntary self-regulation helps us to be independent. We want to be accountable on our own terms.

Another reason for our ambivalence is that we are pulled, simultaneously, toward powerful public policy issues and toward our own issues. The public policy environment — at present strongly influenced by market considerations — is at odds with higher education culture. The important public policy issues today are outcomes, cost and price, consumer demand, competitiveness, and higher education perceived as an "industry." The important accreditation issues continue to focus on process and

capacity and the teaching-learning relationship as an experience rather than as yielding a product.

Yet a third reason for our ambivalence has to do with being evaluated and being valued. Almost nobody wants to be evaluated (accreditation is a form of evaluation). Everybody wants to be valued (accreditation suggests value).

These are the factors that produce ambivalence. To put it another way, higher education is an enterprise that has experienced increasing public regulation for at least five decades. Our history, culture, and background are anti-regulation (or at least reflect limited enthusiasm for regulation). Public regulation is based on the premise that public support should lead to public return on investment. Our anti-regulation stance is based on concepts such as "let the buyer have trust"; "we are the professionals"; "learning is a process and not a product"; and "our return is a community of learning and knowledge development." Is this what the public thinks is return?

What Do We Do About the Ambivalence About Accreditation?

We might try to get people to enjoy criticizing less, but I am not optimistic that this would work. I do have some hope, however, that we can be advocates for the strengths of accreditation. Accreditation is really about guarding certain values and beliefs: the value of general education, faculty intellectual authority and autonomy, collegiality, institutional autonomy, and the benefits of a site-based community of learning. Accreditation is a way of preserving the academic way of life as we have to come to know it. Guarding and preserving our values and beliefs — these are some of the strengths of accreditation.

Strength is not perfection, however; our advocacy needs to be accompanied by commitment to change — focusing energy on identifying and addressing the challenges that face accreditation. We need to change how accreditation does some of its business, with more emphasis on evidence and community standards, greater responsiveness to the changed public policy environment and the changed relationship between society and higher education, and more public communication.

Accreditation is really about guarding certain values and beliefs: the value of general education, faculty intellectual authority and autonomy, collegiality, institutional autonomy, and the benefits of a site-based community of learning. Accreditation is a way of preserving the academic way of life as we have come to know it.

We can also make accreditation more useful in two other ways. We can pay more attention to defining what counts as quality in distance learning. And we can align accreditation review and institutional strategic goals to make accreditation more useful to institutions.

Although the ambivalence is there, we can deal with it by recognizing the strengths of accreditation and acting in areas of needed change. This is a task for faculty and administration. Both roles call for energy and investment in resolving the ambivalence.

Dealing With the Confusion About Quality: Why Is It There?

A major reason for our confusion about quality is language. Whether the language results from confusion of thought as well, I do not know. Neither do I know whether the confusion in language is intentional.

Some examples of language problems

point to our need to change. We use *quality* as an adjective and a noun. We talk about *accreditation, quality assurance,* and *assessment* as if they were the same — they are not. We attach "quality" to everything, and then have the temerity to call it elusive and say that we can't define it.

What Do We Want to Do About the Confusion About Quality?

We can, however be clear about the differences among accreditation, quality assurance, and assessment. *Accreditation* is the particular United States form of quality assurance — it is a set of practices based on a set of values. *Quality assurance* is about defining quality and then finding evidence that it exists according to that definition. *Assessment,* as I understand AAHE's use of the term, is finding out what students know and do; assessment focuses on quality in the teaching and learning enterprise.

CHEA offers one way to deal with confusion about one of the three terms, *quality assurance.*

How CHEA Proposes to Deal With the Confusion About Quality

CHEA's approach to quality focuses on the expected results of institutional efforts. The CHEA approach is to strengthen quality through additional attention to results.

If you were to use the CHEA approach, your institution would

- Acknowledge the value of operational definitions of quality, rather than seeking an ideal and insisting that quality cannot be defined.
- Ensure that institutional mission is central to definition of quality.
- Develop expectations of results for all major institutional activities — teaching (e.g., student learning gains), research (e.g., patents and impact on specific research areas), and service (e.g., evidence of impact on local community).
- Develop expectations of results in the context of institutional resources and the educational profile of students.
- Obtain evidence needed to confirm results, such as performance, process, and resource indicators.
- Evaluate actual results in light of expected results.
- Examine actual results in light of information about results from similar institutions.

Quality is affirmed when we have set expectations for results in light of institutional mission and resources, obtained evidence of results that are achieved, and compared the evidence of results with expectations.

CHEA sees quality assurance through accreditation as an examination of three key dimensions of an institution: *resources, processes,* and *results.* Too much time is spent on the first dimensions of resources and processes, and not enough time is spent on the third key dimension, results. Defining *quality* as "results" means that CHEA will advance quality through particular attention to the results dimension.

We can alleviate confusion about quality by clarifying the use of language and choosing the means by which we focus on results. Alleviating confusion, again, requires both faculty and administrative caring and concern.

The Good News

Ambivalence and confusion are with us. The good news is that we are positioned — if we want to be — to deal with both. The ambivalence can be dealt with by advocating the value of accreditation, changing

how we do some of our business, and strengthening the usefulness of accreditation. The confusion can be dealt with by clarifying language, defining what we mean by *quality,* using this definition consistently, and developing accreditation review practices that produce evidence for quality defined as results.

It is fashionable to bash accreditation. And, as with any important undertaking, accreditation has its strengths and weaknesses. Yes, it needs reform. But it is certainly more desirable than some of the alternatives — such as government regulation and market regulation.

Is there a faculty role in accreditation? Of course. It is to work with other faculty and with administrators to further our defining beliefs in higher education, our commitment to change, and our strengthening quality through attention to results. ●

.....

Judith Eaton currently serves as president of the Council for Higher Education Accreditation, a national organization promoting quality assurance and accountability in higher education through voluntary accreditation.

During her career, Eaton has held positions as state chancellor, president of two community colleges, president of the Council for Aid to Education, and vice president at the American Council on Education. She has also served as a full- and part-time faculty member in several institutions. She has been awarded three honorary degrees. Eaton has published numerous articles in higher education journals and periodicals, edited several publications, and is the author of two books on higher education, The Unfinished Agenda: Higher Education in the 1980s *(American Association of Community Colleges, 1997) and* Strengthening Collegiate Education in Community Colleges *(Jossey-Bass, 1994).*

I see the American Association for Higher Education and the National Academy of Sciences as close allies, both being committed to change, as the name of your major magazine emphasizes. Many people are surprised to learn that the National Academy of Sciences, which sounds like a pretty old and

Reinvigorating Science Education in the U.S.: The Importance of the Appropriate Assessments

by Bruce Alberts

staid organization, is committed to change. Our organization needs this focus because science and technology are rapidly changing our world, and are increasingly becoming the main drivers of our society and our economy. Unfortunately, most institutions are very conservative organizations largely designed to maintain the status quo. Scientific organizations like ours can't be about anything but trying to help these institutions, and the people in them, change and adapt to the new realities.

One change that everybody knows about is that driven by new computers and communications. Since I arrived at the Academy in 1993, I have learned from the experts in the field that we are only at the very beginning of this revolution. Arthur Schlesinger, Jr., a former teacher of mine at Harvard, has said that he believes that the transformation of society that will eventually result will be as profound as the Industrial Revolution was in transforming the world from an agricultural to an industrial society. We know this change is inevitable, and the more we accept the change and exploit it in good ways, the better off we are all going to be.

The problem is a seven-letter word: *inertia*. There is much more inertia in human society than there is in physics. In physics, if you push on something enough,

no matter how heavy, it moves a little bit. Human societies, however, are set up to be stable systems. Time after time, for example, talented and idealistic people try to improve our schools, instituting major projects with major effort; yet when the projects end, the schools slide back to where they were before. It is this inertia that we must all work together to overcome.

Let me now introduce you briefly to the National Academy of Sciences. I would like you to see the Academy as a friendly place that you would like to interact with and visit. We are already a major tourist attraction because we have a famous statue of Einstein on our front lawn — on Constitution Avenue, just across from the Vietnam Veterans Memorial.

The National Academy of Sciences was founded in 1863, when Abraham Lincoln was president. At our inception, we got a special charter from Congress that makes us different from any other organization. In return for the right to exist as a private organization and an honorary society of the nation's best scientists, our government charter requires this of us: "The Academy shall, whenever called upon by any department of the government, investigate, examine, and report upon any subject of science or art." But here's the catch — "The Academy shall receive no compensation whatsoever for any services to the government of the United States."

It is not clear how my predecessors felt about this requirement in the old days, but in retrospect it has been a great advantage because it has caused us to be infused by a volunteer spirit. We enlist the efforts of thousands of volunteers every year, and we are very much a service organization.

Today, the National Academy of Sciences is part of a larger entity. Our operating arm, called the National Research Council, was founded during World War I to bring in volunteers besides scientists — i.e., teachers, lawyers, and engineers — to help give advice to the government. Subsequently, two other academy-like organizations were founded: the National Academy of Engineering and the Institute of Medicine.

The three organizations work together to run the National Research Council, which nearly every working day publishes a report on some subject that the government has asked us about — in total, about 200 reports a year. Most of these reports are available for free to anyone who wants to read them or print them out from our website <http://www.nas.edu>; in addition, bound copies can be purchased directly on the Web.

A large number of the studies that we carry out are, in fact, assessments. We assess governmental programs such as the Partnership for a New Generation of Vehicles or the Environmental Protection Agency's proposed research that will lead to its regulation of airborne particulates. We assess several different governmental research laboratories. We frequently assess the state of scientific knowledge with regard to potential risks to human beings in our society.

For instance, we completed a major report on the health hazards from electromagnetic fields that one encounters in the home from appliances, electric wires, and power lines, which appeared on the front page of nearly every newspaper a few years ago. Looking at data from the many scientific studies conducted over the past two decades, we concluded that there is no evidence that these kinds of electromagnetic fields are dangerous. (This evidence notwithstanding, many people remain frightened about electromagnetic fields, and as a

result, billions of dollars have been spent protecting people from false dangers.)

Assessment as Investigation

Let me turn now, however, to assessments of a special kind: assessments in education. Like you, I have had extensive experience with education: I have been a professor in universities for thirty years, first at Princeton and then at the University of California-San Francisco, and I have spent a lot of time in school systems.

All those who have had these kinds of experiences recognize the enormous power of assessments. Because we tend to get the type of education that we measure, we need to pay much more attention to the exact nature of the many tests that we give to students. While a professor I didn't think enough about this important issue, although I did see some horrible tests.

My first introduction to scientific reasoning was through junior high school and high school geometry, which I loved, so I have constructed a little theorem about assessment: "What is measured in high-stakes assessments has a profound effect on human behavior." I can't emphasize that enough. The corollary, therefore, is, "We must be exceedingly careful to make sure that we measure what counts." Another important point is that if we don't measure it, it may not exist: When we measure some things and not others, that which is not measured tends to get neglected.

I believe that, due to inertia, our system of education with respect to science and math education is broken at nearly every level. Moreover, for the above reason, we cannot expect major improvements in this system without major changes in our assessments both of students and of faculty performance.

Due to inertia, our system of education with respect to science and math education is broken at nearly every level. Moreover, for the above reason, we cannot expect major improvements in this system without major changes in our assessments both of students and of faculty performance.

The Academy took our most comprehensive look at the whole system when the National Research Council was asked to oversee the preparation of the National Science Education Standards.

You might remember that in 1989, the fifty state governors met in Charlottesville — then-Governor Clinton was the leader of that group — and they called for the first-ever national education standards in major academic subjects for kindergarten through twelfth grade. In 1991, after the hot potato bounced around a while, the Academy was assigned the task of preparing the National Science Education Standards. This task took four years, involving successive drafts with extensive public comment, and contributions were made by thousands of teachers, scientists, and science educators.

The net result, a 250-page book (available on the Web at <http://www.nap.edu/readingroom/books/nses/>), drew important conclusions and made many significant recommendations.

To me, these standards have three bottom lines. First, science should be a core subject in every year of school starting in kindergarten. While true in the United Kingdom, this is not the case in the United States, where science is often viewed as an enrichment, like band.

Second, and very important, science must be for all students, not just those who might someday become scientists or engineers. We live in an increasingly scientific and technological age, and acquiring some

fundamental scientific skills will be important for everyone in our society in the twenty-first century.

Last, and most critical, science is not the memorization of all the parts of the flower or the parts of the cell, or becoming familiar with science word definitions and facts. Science education instead should emphasize inquiry-based learning and problem solving, as well as science understanding that can excite children and empower them for the rest of their lives. In this kind of science, classes look different. For example, my favorite classrooms in the San Francisco public schools are noisy, with the students challenging each other and the teacher playing the role of a highly skilled coach, not just standing in front of the class spewing out knowledge to be memorized.

To that end, the Academy and the Smithsonian Institution, through the National Science Resources Center, directed by Douglas Lapp, have produced twenty-four sets of eight-week science modules appropriate for each grade level, one to six. Each module comes with a box full of materials for thirty students, plus detailed instructions to the teacher on how to guide students through learning by doing. This kind of science was introduced as the required curriculum in the San Francisco public schools — with the aid of my university, UC-San Francisco — shortly before I left for Washington.

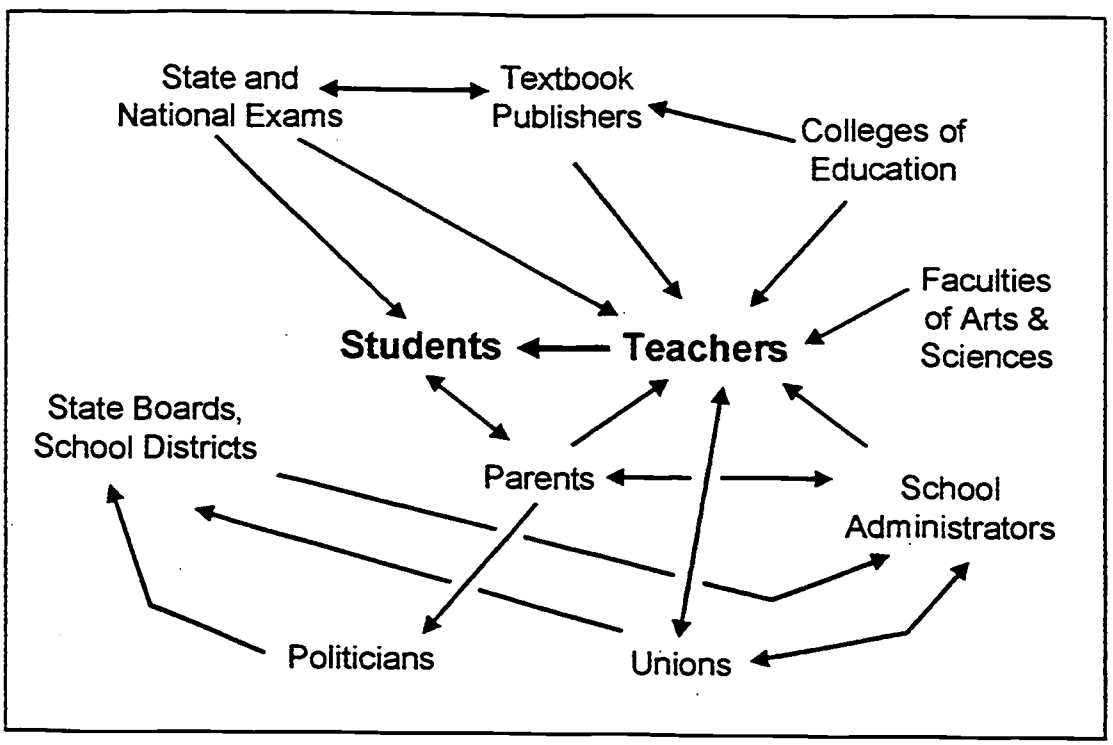
The major message I would like you to remember is embedded in the title of a booklet recently produced for parents by the National Research Council — "Every Child Is a Scientist." This whole country works on slogans; if you say something enough times, people will believe and understand it. So that's what we need to keep on saying, "Every child a scientist." This is what we should aim for in our society, and in this booklet of twenty-six pages (available at <http://www.nap.edu/readingroom/enter2.cgi?0309059860.html>), we explain why.

It is hard to imagine how we can accomplish the aim of making every child a scientist in a system with so much inertia. Chemists would describe our present education system as being in a "stable equilibrium": a stable state in which multiple forces support one another. Such a view of our education system is shown in Figure 1. There are many different players, all of whose actions are critical. You'll notice that most of the arrows are pointing at the teachers, who are very ill-served by this system. Also notice that my own colleagues, the faculty of arts and sciences, are only very poorly connected to most of the important elements.

For the purpose of this talk, I want to focus on our system of state and national examinations and its interaction with textbooks. Our current textbooks teach to the state and national examinations, and the national and state examinations teach to the textbooks. In science, this is why we have all these words to be memorized.

When I asked the Educational Testing Service how it decides what to include in its Biology SAT II exam, I learned that ETS sends out to teachers a mass questionnaire with a list of topics, asking, "Are you teaching this?" When it compiles all the answers, lo and behold, it discovers that its exams are just right: They're covering all the topics and words that the teachers are teaching. It concludes that everything is fine in Testland, in a stable state of equilibrium. Until very recently, the test writers have not thought it important to talk to outstanding teachers and ask them how the system is affecting their lives and their teaching. Without such direct, continual communica-

Figure 1


State and
National Exams
Textbook
Publishers
Colleges of
Education
Faculties
of Arts &
Sciences
Students
Teachers
State Boards,
School Districts
Parents
School
Administrators
Politicians
Unions

We all laugh about this, but this is what millions of our students are being exposed to, and we wonder why they don't value education. If this is what adults think is important, and adults are after all making these tests and making these books, let's just watch MTV and forget about it.

tion with the most critical people in the system, we will never get out of our gridlock.

Let me give you some examples from my personal experience of what's broken at three different levels of our present educational system. Since I'm a biologist, let me start with seventh-grade biology, which is often a true horror. From textbooks about the human body, or the parts of the cell, students are expected to memorize an incomprehensible list of information — reflecting all of the science knowledge we want them to be stuffed with.

Several years ago I was asked by an organization called Textbook Letters to review a very popular, 500-page middle school life science textbook. After I had read all 500 pages, I concluded that this was the hardest book I had ever read, because it really didn't tell you enough about anything to acquire any kind of understanding. The only way to deal with the material therefore was to memorize it, as if it were a vocabulary list in a foreign language. Should we wonder why so many kids in junior high school are turned off by science, by education, and by school?

Permit me a brief quote from this textbook. This is what the book said in the chapter that describes all of the parts of the cell: "Running through the cell is a network of flat channels called the *endoplasmic reticulum.* This organelle manufactures, stores, and transports materials." The next paragraph is about the Golgi apparatus, and the textbook continues like that for pages and pages. I happen to pick this particular quote simply because, sixteen pages later, there is an end-of-chapter self-test, which purportedly emphasizes what is important to know. The self-test asks — I'm quoting, not making this up — "Write a sentence that uses the word *endoplasmic reticulum* correctly." Now how would you feel about an educational system that was making you memorize such meaningless sentences? We're turning middle school kids off from real learning.

Let me move to a higher educational level. Four years later, in the middle of high school, you're going to take your achievement exam in biology — it's now called the SAT II exam. Again, it covers all of biology, with no emphasis on understanding. Let me quote from the 1997 edition of an exam-preparation book called *Cracking the SAT II: Biology Subject Test:*

> *We'll show you that you don't really have to understand anything. You just have to make a couple of simple* associations, *like these. Aerobic respiration with: presence of oxygen more ATP produced.... Anaerobic respiration with: absence of oxygen, less ATP produced.... When we get through, you may not really* understand *much about the difference between aerobic and anaerobic respiration. But you don't have to, and we'll prove it.... Whether or not you understand your answers, the scoring machines at the Educational Testing Service will think you did. Their scoring machines don't look for brilliant scientists and they don't look for understanding.... Stick with us, and you'll make the scoring machines very happy.*

We all laugh about this, but this is what millions of our students are being exposed to, and we wonder why they don't value education. If this is what adults think is important, and adults are after all making these tests and making these books, let's just watch MTV and forget about it.

Finally, let's fast forward another four years. Now the exam consequences are really getting serious. Juniors in college will take a high-stakes test, called the MCAT, for entrance to medical school. We want doctors who can think, not just memorize, but we don't test for thinking and understanding in this exam — just a staggering amount of memorization. For seventeen years I taught part of a first-year cell biology and biochemistry course to 150 medical school students of the University of California-San Francisco. These are some of the very best students in America, since we compete evenly with Harvard for medical students.

When I arrived there in 1976, all of the tests in our course were multiple-choice exams that could be graded by a Scantron. The professors noticed that most of the students really weren't interested in anything we had to say, except for wanting us to be very explicit about what they had to memorize for the exams. When we talked to the students, we realized they — some of the best students in the United States — weren't learning anything for understanding. So then we created a more complicated multiple-choice exam. We made it multiple, multiple-choice. The answer could be "all of the above," "a only," "a and c," "none of the above," and so on. This test construction took an enormous amount of time and was very hard to do. But after we had put in the effort for a couple years, we noticed that the new test format had not made much of a difference.

Finally, we bit the bullet and made half of the exam short essays. This immediately changed the students' whole attitude about what they were supposed to know. All of a sudden they had to understand something. This change amazed me, and it was my first encounter with the real power of tests. How important it is to get the tests right, if we want to get the learning right — and if we want our educational system to function well and our students to value education!

The National Science Education Standards were written by people who recognized that our education system is a complex, stable system in gridlock. Our committees — people from the front lines, volunteers from all around the country — gave the governors more than they had originally wanted. The governors asked for "content" standards: what every student should know about science in fourth grade, eighth grade, and twelfth grade. We gave them that, in about 125 pages of a 250-page document. The other 125 pages describe the many other parts of the system that will need to change if we are going to change the way that teachers teach, and students learn, science. These changes comprise a rather long list.

The table of contents for our Standards is shown in Figure 2 *(on the following page)*.

The teaching standards are in chapter three. If you want to see what teaching is, how challenging it is to do it right, and to become inspired about being a teacher, I encourage you to read those twenty-five pages — my favorite part of the document. There are also standards for teacher professional development — that is, the education and continual updating of teachers. Chapter five is the most appropriate to today's topic of assessment. It concludes with a summary statement that advocates for "less emphasis

Figure 2

SCIENCE EDUCATION STANDARDS

Contents

1. Introduction
2. Principles and Definitions
3. Science Teaching Standards
4. Standards for the Professional Development of Teachers of Science
5. Assessment in Science Education
6. Science Content Standards
7. Science Education Program Standards
8. Science Education System Standards

on assessing what is easily measured" and "more emphasis on assessing what is most highly valued." That's really the whole message of my talk: We need to put less emphasis on assessing scientific knowledge and more emphasis on assessing scientific understanding and reasoning. Today's assessments really do need to change.

Most of the fifty states are now developing their own assessments and their own science standards. Some are quite remarkable. For example, Maryland has developed the Maryland School Assessment Program, which involves a week of testing every year for third, fifth, and eighth graders. Rather than merely compartmentalizing science, mathematics, reading, and writing, they test for multiple abilities at once.

The following question, which appeared in the *Washington Post* a month ago, is one asked of all Maryland third graders. Here's the problem:

> *Your teacher has received a bouquet of flowers and is having trouble with them. The leaves are drooping, and the flowers look sick. You decide to do an investigation to discover what might be wrong with them.*

Students must then perform the following tasks:

> *(1) Read two articles about plants and their stem system. (2) Write an essay explaining how you would study your teacher's flower to determine what's wrong with it. (3) Draw an illustration that would help other students understand your investigation. (4) With a partner, use a magnifying glass, look at the cut edge of a bottom of a celery stalk [which is used in place of the flower], make a list of things you observe about the stalk, break the stalk, and describe what you see. (5) Draw and color a picture of what you think will happen to this celery if it sits in red dye overnight. Explain why you think so. (6) On the next day, study the celery that was soaked overnight in the red dye. Write a paragraph to explain how the celery is the same or different from what you predicted yesterday. (7) Write an essay explaining why a scientist might want to do more than one investigation when trying to answer a question about science.*

And last,

> *Write a note to your teacher telling what you have learned about flowers and how to take care of them.*

Now that is what I would call a good exam, because it tests for the type of abilities that we want kids to acquire to prepare them for the real world. And it makes school clearly meaningful to them. With that kind of question, parents can appreciate the relevancy of school to their children's lives, and see its importance for getting a skilled job. This kind of assessment stands in stark contrast to testing for the memorization of the parts of the cell, or all those other befuddling demands that we're making on kids in most current assessments.

Unfortunately, the Maryland test is unusual, if not unique. Each state is doing its own thing with regard to assessments, and most are not nearly as inventive or interesting. I should also emphasize that the Maryland assessments are written by teachers and graded by teachers over the summer, and thus represent a great professional-development exercise for the teachers.

Thus far I have let most of you off the hook, because I've given you the wrong impression of who is to blame for poor assessments. Having spent thirty years in universities, however, I think if anyone's to

If anyone's to blame for the current state of K-12 science and math education, it's us — the faculty of colleges and universities. We set the standards. If we use multiple-choice exams, everybody else is going to use multiple-choice exams. If we only lecture at students with a bunch of facts about biology, . . . then of course our high schools will emulate us by doing the same thing.

blame for the current state of K-12 science and math education, it's us — the faculty of colleges and universities. We set the standards. If we use multiple-choice exams, everybody else is going to use multiple-choice exams. If we only lecture at students with a bunch of facts about biology, and if we try to cover all of biology in one year so students can't really understand much about anything in particular, then of course our high schools will emulate us by doing the same thing.

The Advanced Placement course given to advanced high school students is modeled after our course, the freshman high school biology course is modeled after that, and even that seventh-grade course in life science often adopts the model. If we professors admit that the MCAT is a stupid exam, but say to ourselves that it doesn't make any difference — or if we claim that the SAT II exam that we're using for college entry is convenient, even if insufficient — then we're causing the problem instead of being part of the answer.

The Academy therefore has a major focus on improving college-level courses. We have established a new Center for Science, Mathematics, and Engineering Education, chaired by Academy member Donald Kennedy, the former president of Stanford University. One goal of this center is to wake up our sleeping colleges and universities about the need for change. Work at the center has resulted in an overview publication called *From Analysis to Action: Undergraduate Education in Science, Mathematics, Engineering, and Technology* (see <http://www.nap.edu/readingroom/books/analysis/>). A report from the National Science Foundation entitled *Shaping the Future* arrives at the same conclusions (<http://www.ehr.nsf.gov/EHR/DUE/documents/review/96139/start.htm>).

In a more action-oriented mode, the Academy has also published a small book to help those who teach college science, called *Science Teaching Reconsidered* (<http://www.nap.edu/readingroom/books/str/>). I recommend it to your faculty members. Here we raise the question of what science teaching should look like at the college level, especially the first-year science courses for majors and non-majors. Spreading best practice is emphasized. Featured on the cover is a photo of a classroom lecture hall that would probably look very strange to you.

The lecturer is using a technique, developed by Eric Mazur at Harvard, that is now spreading around the country. In his large Physics I class, Mazur stops lecturing every fifteen minutes to ask a conceptual question, which he knows that half the class will get wrong. Students raise their hands to indicate their answers. Neighbors inevitably will have different opinions, and the students then try to convince their neighbors that they are right. After a noisy discussion that lasts for two or three minutes, the students vote again. Now, 85% get the answer right. This technique takes advantage of the fact that someone who has just learned something can often explain it

better to someone who doesn't understand it than can the professor, to whom it is obvious. And, most important, the technique keeps the students awake, alert, and motivated during class. Evaluations of student learning in courses that use the technique prove that it really does work.

Another issue that's being attacked both by the Academy and by AAHE is the question of how we should measure faculty performance with regard to teaching. Remember the statement "If you can't measure it, it won't be valued." We can readily measure research productivity through faculty publications, a method that the faculty trusts and therefore values. If we don't measure teaching performance in a way that people trust, then it can't be valued. This is a very serious issue.

The Academy has begun a new project using some of our endowment funds that looks at how we can best evaluate science and math teaching. Marye Anne Fox, a distinguished Academy member and chemist and the new chancellor of North Carolina State University, is chairing this committee. We plan to coordinate our effort with AAHE.

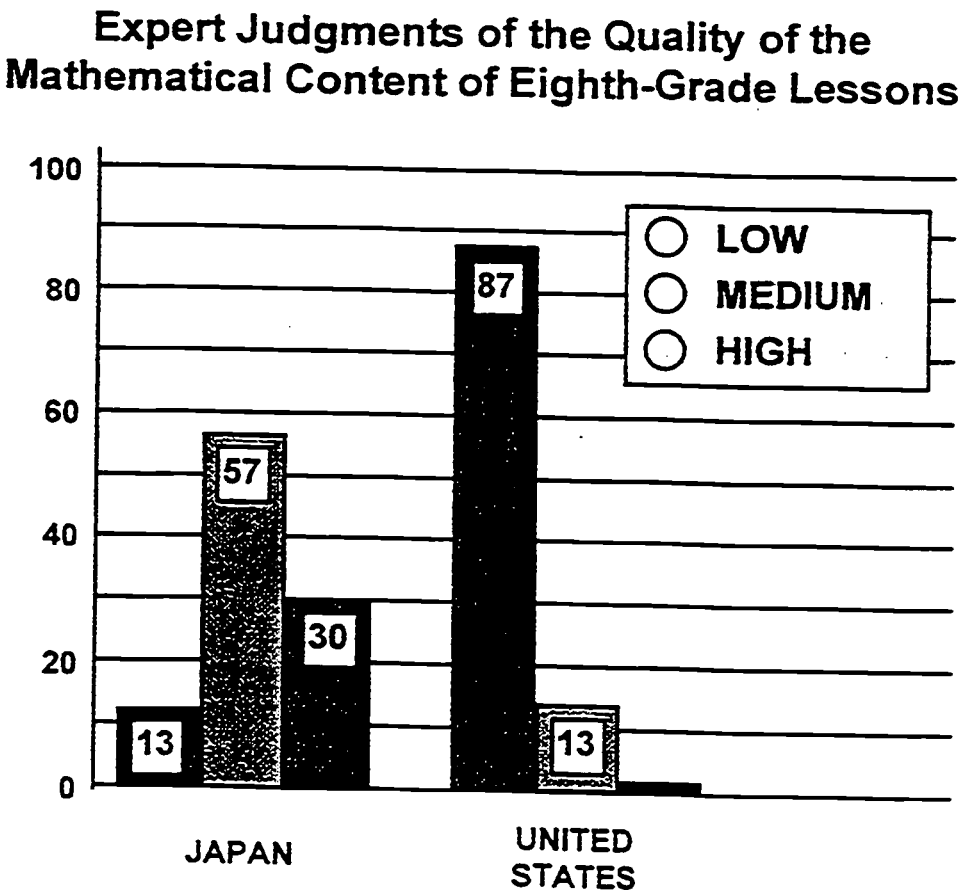
An important international comparison reveals how poorly we've done in science and math education. In February of this year, I had the distinct nonprivilege of helping the Secretary of Education announce that U.S. twelfth graders were basically last in the world in their science and math achievement, according to the Third International Mathematics and Science Study (TIMSS). My friends said, "But the kids in the suburbs, they must be doing really well." But a comparison of the very best students in the United States with the very best students in other countries shows an even worse outcome. In assessments of students taking calculus and advanced physics, the United States didn't beat any nation. The average score for international students in mathematics was 501; ours was 442. For international students in physics, 501; for U.S. students, 423. We weren't even close to the average.

This is an emergency. It is also a wake-up call that we're doing something wrong in this country in education. Americans respond well to emergencies; just think of World War II. So let's start responding.

What should we do first? The TIMSS exam told us something very important about U.S. teachers, because it was accompanied by a very interesting study carried out by Jim Stigler, in which randomly selected eighth-grade math teachers in the United States and in Japan were videotaped. Those videotaped lessons were then graded by experts in math teaching. One of the expert graders said that many of the Japanese lessons were so beautiful that they brought tears to her eyes. Unfortunately, she and her colleagues couldn't say that about the U.S. lessons. The average results are indicative of the problem, as illustrated in Figure 3 *(on the following page)*, in which the quality of the mathematical content of the eighth-grade lessons was ranked as high, medium, or low. Of the Japanese teachers, 30% were ranked high, 57% were medium, 13% were low. Of 100 American teachers, not one was high; only 13% were medium, and 87% of the U.S. eighth-grade math teachers presented low-quality lessons. This difference between Japan and the United States obviously goes a long way toward explaining why their students do so much better in mathematics than ours do.

So we have to ask ourselves, Where do our teachers get taught how to teach? How do they get educated? Teachers are not being well served in our present educational

Figure 3


Expert Judgments of the Quality of the Mathematical Content of Eighth-Grade Lessons
100
80
60
40
20
0
LOW
MEDIUM
HIGH
13
57
30
87
13
JAPAN
UNITED STATES

system. We have to do enormously better. As a trained scientist, I have spent my entire adult life trying to continuously improve my knowledge in science. Scientists build on what other people have done. We take everybody else's advancements, put them together, and then take the next small step forward. That is how science and technology continuously improve.

We need to build the same kind of continuous improvement cycle into our education system, and particularly into the way that we educate teachers. When my wife took her education courses, she found that the professor often required students to purchase the professor's own textbook. These were not good books. I found them almost impossible to read myself, and so did she. Such an education does not contribute to a continuous improvement cycle. Instead, we need to pool our best resources and best practices, if we are to make major improvements in how we teach science and math.

When scientists face a problem such as this one of teacher education, which has been unsolvable for many years, we look for new tools. The new tool that the Academy will focus on in the next year is the World Wide Web, using it as a powerful way of sharing best practices and knowledge. An experimental project that begins this summer [1998] starts with a summer camp for those who have done the best job of preparing middle school mathematics teachers. We're asking these people to bring all their best videotapes, curricula, and class exercises for a show-and-tell. Our aim is to pool the excellent materials from all the best teachers of teachers and put them up on the Web for others to use. Then in January [1999], we'll get several sites around the country to test these materials to see how they work. In the summer of 1999, we'll come back again to improve the website, based on the real-life experiences. Through this small-scale experiment, we will see whether we can begin to make a science out of teacher education.

Japan believes, as many others now do, that teachers can't just get educated in college, then go off and teach. They must have good professional-development opportunities built into their school systems. The Japanese lessons are so good because the teachers keep improving them and talking about them with their fellow teachers. They are given time during the school day for this kind of professional development. In contrast, we seem to assume that teachers can learn everything they need in college, go into a classroom, close the door, and that's it.

Asking "How Did We Do?"

Let me end by looking at another place where we can certainly use continuous improvement. You might think that the colleges that prepare teachers would do something obvious — invite their graduates who have been in the field for two or three years to answer questions such as "What did you learn that was most useful?" and "What didn't you learn that you needed?" These colleges would then change the curriculum every year to improve it, based on this feedback from practicing teachers. Now I've been looking. If there is an education school that is doing this, please let me know, because I haven't found any.

The only place where I have found this kind of continuous improvement process in place is in a program called Teach for America, invented about ten years ago by a Princeton undergraduate, Wendy Kopp. Teach for America places recent college graduates into some of the most desperate schools in America.

These graduates have science, math, or English degrees with no education courses. For five or six weeks during the summer, they go to a boot camp to prepare them to be teachers in these very difficult schools. As an advisory board member for the science and math division, I've attended this boot camp. It meets all day through the evening, six days a week. Although it's not enough preparation, these dedicated students make it work. The program is taking in 700 new teachers this year.

The leaders of Teach for America are doing what every education school should be doing: They're calling in their teachers and asking them what preparation they didn't get that they needed, what was good and bad about the summer institute based on what they now know as experienced teachers? Through this feedback loop they have been continuously improving their preparation programs.

The young people who are running this program have been severely criticized by much of the education establishment for bypassing the normal teacher-credentialing process. But the school principals rank the young Teach for America teachers in the top 25% of all of the teachers in their schools. There's not something wrong with the program led by these young people; there's something wrong with the rest of the system.

We need to take the whole education issue much more seriously. The future of this country depends most of all on what is now called "human capital." As we see from the TIMSS exam, we are not developing human capital. Instead, we're living off our past. If we do not do better in the next twenty or forty years, we will no longer be — can no longer be — the leading nation in the world.

To take this issue seriously means that the most talented and able people in this country have to pay attention to it. I'm very pleased that AAHE is so deeply involved. I hope you will agree that the Academy has been trying to do our part, but we need many more players. We need your universities to get involved. We need all of our major institutions in this country to contribute.

I want to end with a quote, my favorite quote about education, from Alfred North Whitehead. It sums up what I've been saying about this whole enterprise:

> *The art of education is never easy. To surmount its difficulties, especially those of elementary education, is a task worthy of the highest genius. But when one considers the importance of this question of the education of a nation's young, the broken lies, the defeated hopes, the national failures which result from the frivolous inertia with which it is treated, it is difficult to restrain within oneself a savage rage. In the conditions of modern life, the rule is absolute. A country that does not value trained intelligence is doomed.* ●

.....

About himself, Bruce Alberts, president of the National Academy of Sciences, says:

"Many different experiences have led me to believe that science education must be transformed to look like science as it's actually practiced. Scientists don't sit around and

memorize lists of obscure terms and science facts. Nor do they follow rigid recipes in their laboratory work, so that science becomes indistinguishable from cooking. Neither should science students. Science classrooms should reflect the real world of science, with individuals working in teams, testing ideas and new approaches, striving vigorously to figure out why things are the way they are. This is what leads to new knowledge and scientific discovery.

"Over the many years that I taught medical students at UC-San Francisco, I came to realize that while they were terrific at memorizing terms so that they could perform well on standardized tests, in the end most of them had little in-depth understanding of the science. When several of us on the faculty retooled the tests so that, instead, these students had to answer questions with essays, the results were astounding: Suddenly they realized they had to really understand rather than memorize.

"Tests and textbooks must be reworked in ways that promote real understanding. I learned a great deal from writing my own textbook: that it's absolutely critical that texts be written in a way that challenges the student to think analytically and not simply be capable of regurgitating a list of memorized terms. Science words are not science.

"Finally, I learned an enormous amount from my daughter, who has been a high school science teacher in the California public schools. She helped me understand the great challenges of teaching today; teachers are under enormous pressure and get very little support. If science education in America is truly to be transformed, then it must begin with a transformation in the way we prepare teachers and a commitment to support their professional development throughout their careers. Teachers are at the center of the education process; in fact, the success of the entire system depends on their ability to engage and harness the intellectual potential of their students. When the education they receive and the school systems they work in constrain them from being able to do that, then we all lose."

American Association for Higher Education
One Dupont Circle, Suite 360
Washington, DC 20036-1110
ph. (202) 293-6440
fax (202) 293-0073
www.aahe.org

ISBN 1-56377-044-X

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").