

Perpetual Movement: Assessment After Twenty Years

Peter Ewell

National Center for Higher Education Management Systems (NCHEMS)

Through a combination of personal choice and considerable luck, I have been in the middle of what has come to be known as the “assessment movement” for some twenty years now. In fact, I recall vividly being behind the podium for a keynote at the very first national conference on this topic that AAHE sponsored back in 1985. Since then, I have had a peculiar itinerant history with assessment—writing, visiting campuses to help where I can, and watching you do your work. The opportunities this experience has given me have been enormous and I have plundered the work of all of you many times over the years for ideas that I can share with others. So I thank you sincerely for all you have done.

I recently had an occasion to think more systematically about how this field has unfolded when my friend and colleague, Trudy Banta, asked me to prepare a brief “history” of assessment as the lead chapter in a just-published book on *Building a Scholarship of Assessment*. As I began doing so—trying to marshal lost skills as a political scientist/historian—I was first struck by the fact that assessment *has* a history. And with this came the realization that I had grown very used to thinking of this business as something novel and faintly revolutionary. It is not a long history, to be sure. But it is one, which like any history, has had its changes of fortune, its shifts in political landscape, its advances in technique, and the constant appearance of new and colorful characters on the scene.

As I got into this task, I was consistently struck by a central conundrum occasioned by the frequently-asked-question, both then and now, “When is assessment going to go away?” Most

things called “movements” do go away. In fact, they suffer one of two typical fates. Either they disappear as “fads,” as TQM arguably did as a sub-plot within assessment. Or they equally disappear as distinct phenomena because their practices become institutionalized and their core values are embedded in the wider culture. So far, the assessment “movement” has experienced neither. On the one hand, look around you. Levels of activity are unprecedented, and very few in higher education are at this point claiming that assessment is a fad that will quickly fade. On the other hand, there are equally apparent indicators that despite two decades of activity, assessment has yet to become a “culture of use” on most campuses, embedded in the lives and work of those who inhabit them. The result is the curious condition that inspired my title for this morning—perpetual movement—caught as in the myth of Sisyphus in a task that is never complete, but unable to escape from the task itself.

The central reason for this predicament was pretty clear to me in what I ended up writing. This was the role of external bodies in promoting, mandating, or merely providing an occasion for assessment over the last twenty years. It’s important to remember that the assessment movement has served two very different (and at times incompatible) agendas from the outset. One is an institutional-change agenda, owned by the academy, that can trace its roots to the curricular reform movements of the early 1980s. This strand of activity is often symbolized by the national reports that provided its intellectual foundations—*Involvement in Learning* and AACU’s *Integrity in the College Curriculum*, both issued in the mid-1980s. But the other is an “accountability for results” agenda, owned by others, that can trace its roots back equally far, and whose advent was also signaled by a major national report—*Time for Results*—issued by the National Governors Association. The behavioral result of this dual agenda has always been visible in this conference. AAHE’s first national assessment conference, held in Columbia,

South Carolina in 1985, attracted a large number of people who were sincerely interested in assessment's promise to improve undergraduate teaching and learning. But an equally large number were drawn by the need to respond to emerging state mandates. There are three times as many of you attending this, the eighteenth, AAHE national conference on assessment. And I suspect that many of you embody a similar mix of emotions. As faculty, you care about teaching and learning, and have a heartfelt desire to improve it. Most of you are probably sincerely taken by (or at least curious about) assessment's potential to help you do so. But the immediate, palpable, proximate reason why you are here is North Central, Middle States, SACS, or WASC.

So I came to what I thought was the quite natural conclusion that the continuing interplay between these two basically different agendas was ultimately responsible for assessment's status as a "perpetual movement." The persistent requests of public and "semi-public" actors like governments and accreditation agencies provide an unavoidable stimulus for getting assessment started and keeping it moving. So long as these actors are present and pushing, something called "assessment" will surely be done. But the fact that it almost always begins at the behest of somebody else creates a pervasive feeling on the part of faculty of being "done to" rather than doing. Feelings like this lead to widespread perceptions of assessment as an "administrative" rather than an academic pursuit. At the same time, they produce the understandable reaction of trying to "second guess" what external actors "really want," and a significant reduction in risk-taking.

Hence the core dilemma. On the one hand, assessment can't be dropped because it has become an environmentally-imposed condition of doing business; institutions can no more do without assessment than they can do without a development office. On the other hand, assessment can only rarely be creatively engaged by institutions for exactly the same reasons. This conclusion,

in itself, should stimulate our thinking. But despite its simplicity and apparent veracity at one level of analysis, some recent personal experiences in both the institutional and policy worlds have prompted me to dig deeper into our current condition. Let me share three vignettes from personal experience, all from the past three months:

- Asked to review campus assessment activities across all campuses in a state system, I experienced the curious sensation of going back in time. Here were a set of documents that described processes that were workmanlike and probably workable, but that were virtually identical to their counterparts I had reviewed in a neighboring state in 1986. They contained the requisite committees, matrices, and methods—but showed little use of the kinds of authentic and engaged methods typical of sessions at this conference. When I raised such methods in discussion with campus representatives, most reacted positively and enthusiastically. But always with a questioning note of incredulity—“you mean they will actually let us do that?”
- At a presentation on assessment to the assembled faculty of one of the few campuses left untouched by assessment, I was asked the second-most-frequently-asked-question about assessment by a lively and intelligently-engaged faculty audience: “Is there any real evidence that assessment actually promotes better learning?” (The most prevalent FAQ in assessment, of course, is, “Why do we have to do this?”) In response, I gave the usual answer: some clear instances of success (but admittedly few), most campuses are still getting started, and attributing actual learning gains solely to the previous measurement of student achievement involves a pretty long causal chain. All of these, of course, are right answers. But I found myself very unsatisfied, in retrospect. After twenty years we ought to have an answer to this basic question.

- At a gathering of state officials on the popular topic of performance funding, the presenters were asked what amounted to the same question from a *Chronicle of Higher Education* reporter: “What evidence do you have that this policy approach actually works, that it influences institutional behavior in the ways you want it to?” I found myself replying to this question (and being only partially misquoted in the resulting article) that it didn’t really *matter* whether it worked or not because policymakers, for whatever reason, had already decided to do it. The point was, therefore, to try to make it better—not only so that there was some ultimate chance of doing good, but to improve the inevitable. And it struck me immediately that this comment could apply equally to assessment.

These quite recent experiences prompted me to try to conceptually explore the place we collectively find ourselves in more fully as a social and intellectual phenomenon. And as I began to unpack this situation analytically, I found not just one conundrum, but a complexly-layered set of interlocking contradictions that are deeply embedded in what we do in assessment. I would like to review these with you in the form of four deceptively-simple “dilemmas of practice.” Each of these is fundamentally dialectical in nature. That is, as posed, they *have* no answer, unless we explode our current way of thinking. But precisely *as* dialectics, they can certainly be recognized and made conscious in our practice, can perhaps be managed, and just possibly can be forced eventually to yield a higher resolution. They are: the dilemma of *purpose*, the dilemma of *stance*, the dilemma of *technique*, and the dilemma of *consequences*.

The Dilemma of Purpose. The first of these dilemmas is familiar, and is perhaps the easiest of the four to describe. It has become commonplace in discussions of assessment over the past two decades, for example, to make the distinction between “accountability” and “improvement”—the

first ascribed to “them” and the second to “us.” And the situation gets complicated because, with the possible exception of an occasional highly-caricatured Governor or state legislator, almost everybody *says* they are actually for improvement. The real distinction here, though, lies in the classic difference in purpose captured by the professional evaluator’s distinction between “summative” and “formative”—or, to put it in more colloquial English, the difference between making a decisive judgment about a program or performance and engaging in continuous improvement.

It is important to note in passing here that this dialectic is also deeply embedded in the process of teaching and learning. On the one hand, a good teacher is supposed to model the role of “coach” by giving continuous, non-judgmental, feedback to the learner. On the other (at least in this country) she or he also must periodically grade student performance for the record. The resulting role conflict has been blamed for everything from ritualized student behavior to grade inflation.

And it is equally important to recognize that this is not just a distinction between “external” and “internal” actors. Indeed, in my case-study work with many institutions, I have observed consistently that the appellation “them” begins about two levels up in any hierarchy. For the institution’s President, “them” is indeed the governor or state legislature. But for line faculty, the world of “them” usually is populated by Deans and other internal administrators.

Finally, it is useful to acknowledge the supreme irony that most of those perceived to be demanding “summative” assessment really and sincerely see their role as “formative.” Consider the following commonly-expressed directive: “All we care about is that you creatively and actively examine your own practices in your own way, and start building a culture of evidence.” Based on my experience with many external stakeholders over the years, these initial sentiments

are sincere. But when the response from institutions is ritualized and minimal, these actors become frustrated and simply push harder. This was certainly the logic of state action with regard to assessment in the late 1980s. And it is surely the core dilemma of accreditation and assessment today.

What might we do to begin to address the terms of this core dilemma? A first important breakthrough in our thinking would be to recognize that *both* periodic judgment and continuous feedback are important in occasioning institutional learning, but that (like grading) it is probably counterproductive to try to accommodate both in the same process. For actors like accreditors and quality-assurance agencies, the message is clear. Try to visibly separate the functions of compliance and deep engagement in order to get matters of accountability off the table quickly so that we can move on to inquiry. To their great credit, virtually all regional accrediting agencies are moving in this direction, though the move is admittedly difficult to make.

A second important breakthrough is to recognize and harness the hidden information for “formative” purposes generated by intentionally “summative” situations. Much of assessment practice, for example, involves dreaming up new tools to apply to the process of teaching and learning. But a lot of the information we seek may already be contained in *existing* pieces of student work like examinations, writing assignments, and problem sets, if only we knew how to get at it. All of this work, of course, is “summative” from the student’s point of view—and ours as well, when we credential it for the record. Yet grades as we know them can’t give us the information we are looking for. Techniques like Primary Trait Analysis, though, provide us with a way to unpack and disaggregate the hidden information already contained in student work to reveal patterns of strength and weakness across students on different *dimensions* of ability (for instance, problem formulation vs. argument in critical thinking), as well as patterns of strength

and weakness across different *kinds* of students (for example, demographic descriptors, extent of prior coursework, or learning style).

Another of my favorite examples of needlessly “summative” thinking is at a different level of analysis. Many campus assessment plans begin by establishing goals for learning—a reasonable thing to do—but then go on to establish a “target” level of performance against these goals. So when reporting results, they simply array the available data to determine whether or not this established target was met or not. Under such circumstances, one point over the target is victory, while one point below is disaster. I am continually puzzled by such an approach to disseminating and analyzing assessment results because it deliberately throws away information to achieve an external perception of rigor. Of course it is fine to set targets and see if they can be attained. But we must recognize that behind every target achieved (or behind any central-tendency measure for that matter) is a wide *range* of performance that needs to be looked at. There are certainly plenty of individual cases that are well below where we want them to be, and improving *everybody’s* performance—not just moving the average—is what we should be shooting for.

The Dilemma of Stance. I want to turn now to the second principal contradiction I see in assessment practice, which is a bit more subtle. For a variety of historical reasons, the predominant language of assessment has been that of educational research and the social sciences (note, for instance, my use of the terms “formative” and “summative” a moment ago). These fields are not always the most respected on many campuses and, even if this were not the case, the terminology they promulgate can be both arcane and mechanistic. Partly as a result, assessment *argot* has never sat well with many academics.

But beyond language, the predominant paradigm of about how to actually do assessment has some pretty clear implied prescriptions. First, the stance or position of the observer is quite consciously placed *outside* the phenomenon being observed in order to preserve distance and “objectivity.” Second, the tools and methods employed by this externally-positioned observer are self-consciously “scientific”—quasi-experimental designs, systematic observation, and “testing” proposed hypotheses. Third and finally, the rhythm of observation is punctuated—consisting of a sequence of steps embracing observation and data collection, analysis of the data collected, and testing conclusions against proposed hypotheses that could, in principle, be written up formally as a research report. Though somewhat caricatured in this description, I think these attributes capture a presumed “ideal type” for assessment practice that many campuses, consciously or unconsciously, seek to follow.

It is again important to begin by stressing that this “scientific” stance is not just characteristic of “summatively-positioned” outsiders. Indeed, individual faculty members adopt just this stance when they engage in techniques like classroom assessment—stepping out of their immediate roles for a moment to collect data useful to them for entirely “formative” purposes. And it is equally important to note that this is not just a distinction between quantitative and qualitative approaches to evidence. Indeed, the majority of mainstream evaluation work is now done with qualitative tools like interviews and participant-observation, though the stance of the evaluator remains divorced deliberately and decisively from what’s being investigated.

This stand-outside-the-phenomenon approach is quite different from one in which the stance of the observer is placed squarely and consciously *inside* the phenomenon being observed—in this case, the process of teaching and learning. In contrast to the “scientific,” this is a kind of “craft-based” approach, which relies upon a constantly growing store of tacit, local knowledge of what

works and what doesn't drawn from ongoing practice. Perhaps the most vivid examples of this contrasting stance—and its power—are drawn by John Seely Brown and Paul Duguid in *The Social Life of Information*. Its consequences are also visible in what James Scott labels *metis*, or local practice, in *Seeing Like a State*—a brilliant diagnosis of what goes wrong in large-scale social engineering projects when they defy such knowledge (which should be required reading for governors and state legislators). Such descriptions also emphasize that this contrasting stance is almost always *collective*. Rather than being amassed bit by bit from a single “objective” standpoint, knowledge is instead embodied in a living “folk” setting and is holistic in character. It is by definition context-specific and people-dependent.

I experienced directly how this dilemma can be played out in practice when I served as the evaluator last fall for the new accreditation process being developed by the Teacher Education Accreditation Council (TEAC), an alternative accreditor in teacher education that is currently seeking federal recognition. The TEAC approach to accreditation self-study and review is extremely interesting from the standpoint of assessment because it deliberately treats evaluation as an act of scholarship. Faculty in the program under review are asked to prepare an “inquiry brief” about the effectiveness of their program along the model of a research monograph in educational research. Language or method is not presumed to be a problem here, as it would be in philosophy or physics: after all, as educational researchers, faculty in teacher education programs can be assumed to be familiar with the required terms and methods. And the review process itself is admirably objective. It requires both the faculty and the review team to treat claims about learning as research hypotheses, subject to test using the best methodology of the field. But as it turned out, this approach, in itself, unduly constrained the inquiry about quality because it provided insufficient opportunities for faculty to communicate what they *really knew*

about their programs through years of practice-based observation. The “scientific” stance was important and provided needed rigor. But documenting tacit knowledge was also important.

Such stories can be told about a lot of departmental adventures with assessment. When they depart from the work, as those who do the work understand the work, problems arise. But an additional and more consequential implication of the predominant outside-the-phenomenon stance of assessment practice is that it *adds cost*. This is because assessment must be implemented in addition to, and consciously alongside, the ongoing process of teaching and learning—requiring specially-constructed instruments, assessment days, and special reports. And these parallel activities appended to the regular cycle of teaching and learning tend to reinforce incipient faculty feelings that “this can’t be about us.”

How might we address these defects in the predominant stance for assessment without losing the power and rigor of the classic “scientific” approach? One clear implication here is a trend I see happening already, and of which there are many examples here at this conference. Find ways to engineer assessment opportunities directly into the curriculum at periodic intervals, so that they can do “double duty” as graded course assignments and as opportunities to generate meaningful information about collective performance. Examples include capstone courses, specially-constructed examinations and assignments designed to elicit particular cross-cutting abilities in different content contexts, and work-samples drawn cross-sectionally from a range of classes or gathered longitudinally in a portfolio.

Such “authentic” approaches, of course, have been around for many years, and a lot of folks feel good about them as alternatives to what they see as more mechanistic methodologies. But they will only work well as *assessments* if a number of important conditions, drawn explicitly from

the evaluative/scientific stance to investigation, are met at the same time. First, there must be up-front, concrete agreement about the abilities being fostered, together with clear exemplars of what “success” is supposed to look like. Put another way, “tacit” or “folk” knowledge about outcomes must be disciplined by clear *criteria* of achievement, not just relative ones. Second, there must be well-designed tools to achieve consistency in observations of performance that can be effectively applied *within* a given community of judgment. Too often in otherwise admirable “competency-based” curricula, faculty do not have such tools, and go off on their own to proclaim what’s competent and what isn’t. Third and finally, there must be a collective view of knowledge-utilization in which the community actively makes meaning of the results in the form of retreats, forums, or other collaborative venues. These are hard conditions to achieve, and it is no wonder that we don’t see large numbers of good examples. But they are the principal ways I see to begin to address the dilemma of stance.

The Dilemma of Technique. This discussion leads directly to the third dilemma I’d like to discuss. In fact, it is in many ways related. But I feel that it is useful to break it out as a separate topic because, despite our rhetoric about principles and purposes, a lot of what we actually end up doing in assessment is more “tool-driven” than it is “inquiry-driven.” It has become fashionable to quote Kaplan’s famous “Law of the Instrument” here—which can be roughly paraphrased as the fact that with a hammer in hand, practically everything needs pounding. But it also possible to observe in hindsight wave after wave of what can only be termed “methodological fashions” in assessment over the years—“value-added,” general education testing, portfolios, and now rubrics. All of these are good techniques, if implemented well and rigorously. But all of them also tend to be adopted as a panacea by many, resulting in pale imitations of what the designers really had it mind—“rubric-lite,” if one had to put a label on it.

The primary dilemma of technique, though, has remained fairly constant over the years. The first pole of the contradiction draws heavily on the scientific paradigm by insisting upon constructed methods, validated through the canons of educational measurement practice, that yield largely numeric results. The epitome here, of course, is the large-scale standardized test. The second pole, in contrast, rests heavily upon the notion of “expert judgment” based on years of intimate and successful practice involving viewing roughly similar things time after time. Its epitome, to choose one example, is a juried artistic performance. This never-ending, and at root *ideological* conflict has been played out since the beginning of the assessment movement.

One theme here is the quest for the “perfect” instrument that will simultaneously attain authenticity, validity and reliability. This holy grail of large-scale assessment has been sought for many times. The state-based College Outcomes Evaluation (COEP) program in New Jersey in the late 1980s provides one of the best examples, and resulted in an instrument still in use—the ETS *Tasks for Critical Thinking*. Another example was a series of federally-sponsored efforts in 1990-94 to develop a battery of state-of-the-art instruments to measure the goals of critical thinking, communication, and problem-solving proclaimed a few years earlier by the National Education Goals Panel—an effort that ultimately resulted in an RFP for a test but which went unfunded. Its latest manifestation is a “Value-Added” project currently being undertaken by the RAND Corporation’s Council on Aid to Education (CAE), which was written up in the latest issue of AACU’s publication *Peer Review*. These efforts are all intriguing, but they are very hard to pull off because of formidable problems of cost, institutional buy-in, and student motivation. Because of this, they may be ultimately as useful to faculty as illustrations about how to build a good test as for any actual data they may yield.

But an equally-present and unfortunate *leitmotif* of much of the assessment movement is mindless test-bashing in the name of so-called authentic work. This is also an ideology, that if unexamined can cause us considerable harm. There are plenty of reasonable uses for well-designed standardized instruments both on and off-campus, and we should be mindful of such appropriate applications as well as the limits of use. Much good work has been done here already in extensions of the classic concept of validity by the late Sam Messick of ETS, and in notions of “contextual validity” developed by Marcie Mentkowski and Glen Rogers of Alverno College.

So there are some clear ways to begin to manage this dilemma. A first mindshift that’s needed, of course, is to recognize that in some sense *every* assessment is a matter of expert judgment. Tests are made by people, and they chose what they chose to put on them for a reason. It’s important to understand what that reason was and how close it is to your own purposes. A second needed recognition is that large-scale assessment of higher-order collegiate abilities can be an important tool of social policy, if it is directed toward determining gaps in the nation’s store of “educational capital” and filling them. It may be equally useful in benchmarking progress for institutions by occasionally validating their own local assessment programs from an outside standpoint. But it is neither an appropriate, nor even a very effective, lever for rating and governing institutional performance—as currently demonstrated by the increasingly nasty testing movement in K-12.

We also need to recognize the growing power of technology to overcome some elements of the current contradiction between standardized and authentic methods. If done well, technology-based assessments have the potential to support a seamless and continuous approach to examining student performance through complex simulations and problem-based settings that

allow students to directly manipulate multi-faceted situations and observe the results.

Technology also allows us to track student progress through such problems completely and unobtrusively, noting the paths they take and recording their false starts and diversions. Such approaches promise an eventual fusion between authentic and standardized methods, enabling expert judgment to function on a very large scale.

The Dilemma of Consequences. The final dilemma I'd like to explore returns us in many ways to the beginning. Much of the difficulty encountered when the topic of assessment is first raised at an institution is due to fear of consequences. In fact, the distinction between formative and summative motives I discussed earlier in the dilemma of purpose is often confounded with consequentiality. The question posed is always, "What will happen if they (always an alien) find out X (where X is always something bad)?" The result of this line of thinking is usually a radical reduction in risk-taking, regardless of whether the source of the stimulus is internal or external.

But it is important to view consequentiality from multiple perspectives, and not conflate choices about consequences with choices about basic purposes. One of my favorite examples here is what I have come to call "the paranoia shift," and I have seen it manifest among faculty on many campuses. The first stage of paranoia is familiar: "They are going to use this to get me." The result is often a well-intentioned effort at reassurance on the part of those responsible for assessment: assessment results will not be used in faculty evaluation, won't count for students, are experimental and exploratory, etc. Then the second paranoia kicks in: "Nobody is going to use any of this stuff for anything important anyway, so why should I waste my time?" This common dynamic illustrates the difficulties raised by simple "solutions" to the consequentiality problem.

Instead, consequentiality needs to be carefully balanced to fit both the cultural circumstances of a given institution and the intent that the assessment effort is trying to serve. Such a careful balance of challenge and consequence is, of course, integral to good teaching and learning situations. For example, Art Chickering long ago wrote about the fact that some of the best learning takes place when the level of challenge is placed at a high level, but just within reach. And again, the perceptions of John Seely Brown and Paul Duguid are *apropos*. Writing about the power of what they term “cognitive apprenticeship,” they emphasize the fact that good learning occurs in such situations because the supporting tasks played by an apprentice, though small, are *real* and therefore must be done right.

These same lessons about carefully managing consequences are true at the organizational level. I have seen some of the most effective examples of this in the corporate world—especially in Baldrige-award-winning companies. A few years ago, we partnered with some of these in a project on the effective use of performance measures in organizational management through the good offices of the American Productivity and Quality Council (APQC)—a project in which some of you in the audience may have been involved. The lessons of this encounter about managing consequences were striking, and quite different from the ways we often use assessment information in colleges and universities.

A first lesson is don’t hold people accountable for things they cannot do. We need to recognize the often quite different contexts in which people (or academic departments or institutions) do their work that may substantially affect outcomes. The classic violation of this in accountability policy is the use of graduation rates to judge institutional performance without taking into account substantial variations in selectivity, which is the single largest predictor of graduation. Good practice in management, in contrast, emphasizes tailoring expectations directly and

individually to circumstances through the use of mechanisms like “stretch goals” and monitoring improvement rather than comparative performance.

A second lesson is don’t just automatically reward “good” performance. The knee-jerk assumption for many people who first enter an evaluative situation is that poor assessment results will lead inevitably to negative consequences. The result is either outright resistance or a numbers game in which the object is to focus on manipulating results to look better next time. But results that are below expected levels ought instead to lead to greater investments in order to fix the problem, if the goals involved really are important. This is a lesson that I’ve seen Provosts and Presidents use to their advantage, though not as often as I’d like. And it changes the whole dynamic of the campus conversation, because people are less inclined to hide negative results.

A third lesson is use the resulting information to emphasize collective and collaborative benefits instead of promoting competition among units and individuals. Focusing assessment results on addressing visible problems in the campus environment can be a win for multiple parties, and this should be a focus of our information-utilization efforts. This is why I like prerequisite course sequences so much as a starting point for assessment. Improving them addresses a commonly-expressed faculty frustration—the need to continually “re-teach” material that was presumed to be learned in a prior course but that is demonstrably absent in current student performance. Careful analysis of such situations can reveal opportunities for consequential changes that have a collective and visible benefit. Similarly, in state-level assessment practice, it would make a big difference if the emphasis were placed on determining with considerable precision particular areas of statewide deficiency for *all* institutions with respect to particular

populations—a challenge that they can address as part of the same team—rather than inducing counterproductive competition among institutions by constantly comparing their performances.

These are all good lessons for us when we engage the delicate task of using assessment results to leverage change. And none of them accept the terms of the dilemma of consequences as originally posed: either react negatively to bad news (and scare people away from taking any risks) or have no consequences at all (and watch them spend their time elsewhere).

Returning to Our Roots. Simultaneous consideration of these four dilemmas of practice compels us to recognize the highly nuanced character of assessment as both a practical and an intellectual task. It forces us to make conscious, considered choices about what to do in a given context. Such consideration helps us avoid the kind of simplistic (and frankly ideological) solutions that keep bedeviling our practice.

For example, one particular combination of choices clearly does produce the option of large-scale standardized testing, imposed from the outside, with substantial consequences for poor-performing institutions as is currently popular in the K-12 arena. But this does not mean that other approaches that employ the technology of large-scale testing for different purposes and with different consequences are all bad. For example, recent proposals to look comprehensively at the nation’s “educational capital” at higher levels through a collegiate counterpart to the highly-respected National Assessment of Educational Progress (NAEP) may be good public policy. It’s not standardized testing that’s the culprit here, but rather the particular combination of motive, stance, and consequence surrounding it that we need to avoid.

Closer to home, trying to move beyond the terms of the four dilemmas as posed allows us plenty of room to work on significant alternatives to what has become the prevailing modal type for a

campus assessment program—one that is nominally formative (but looks pretty summative), operates largely as an add-on to other institutional activities in teaching and learning, and has few consequences for anybody involved. The best of these alternatives return assessment to its original roots in curricular reform in the 1980s, a time when it emerged together with some equally powerful reform initiatives like collaborative learning, learning communities, service learning, and problem-based learning. More recently, additional opportunities have arisen with which assessment can partner in re-shaping the landscape of undergraduate instruction. One of these is the growing force of learning-centered pedagogy that demand constant attention to outcomes by individual faculty. Another is the growing salience of competency-based designs for learning that are founded fundamentally on the assessment of mastery. A final opportunity lies in the growing use of technology in instructional delivery that both allows assessment to be seamlessly embedded in the teaching/learning process, and that demands proofs through assessment of its relative effectiveness.

The role of assessment in all of these is far more fundamental than in the straightforward evaluative paradigm. First, they begin with agreed-upon standards for learning, stated in outcomes terms, that *all* students are expected to master. Second, they use the best of assessment technology to certify and credential individual attainment, embedded at strategic points throughout the curriculum. Finally, they harness the assembled results of these assessments to periodically draw conclusions about what needs to be changed. Assessment is not added on to teaching and learning in such conceptions. It is an integral part of it.

I see many prototypes of this approach emerging piecemeal, particularly at this conference.

They include capstone courses and culminating demonstrations that provide excellent settings for simulations and problem-based assessments, posed within a particular disciplinary context. They

include work with embedded course assignments carefully constructed to examine a particular ability in addition to testing content, examined through agreed-upon scoring rubrics. They include asynchronous, mastery-based programs with periodic credentialing assessments, delivered largely through technology. And, at a different level of analysis, they include the beginnings of ability-based articulation agreements among groups of institutions designed to facilitate effective transfer.

I see also the organizational structures needed to support these developments starting to emerge. One of the most important is the establishment of teaching and learning centers at many institutions—places where assessment can be fused with faculty development and other important change initiatives across a single campus. Another welcome development is the continuing progress of the scholarship of teaching that provides a way for faculty efforts in assessment to be rewarded and disseminated. And a final good sign is the growing activity of individual disciplinary associations to both promote assessment within their disciplines and to provide an outlet for assessment scholarship in the form of publications and conference presentations.

But these many elements—both in teaching and learning, and in the necessary organizational support structures—have yet to be adequately integrated in the form of a complete teaching/learning/assessment *system*. Doing so would provide a resolution of the core dilemmas of practice that have stalked our history. With regard to *purpose*, it yields an approach that is simultaneously formative and summative, but at different levels of analysis. For students, assessments are real and certify attainment. For us, the resulting work is sampled to yield useful information for improvement. With regard to *stance* and *technique*, it revives prominently the notion of faculty as a professional community of practice because most assessment will rely on

collective judgment. But such consensus is disciplined by the tools of science to achieve consistency, and is periodically benchmarked against external standards established by cutting-edge assessment technology to assure that standards are met. With regard to *consequences*, the process is consequential for all concerned. For students, performance is meaningful and results count. For faculty, results directly inform the work on a daily basis, and can be directed toward palpable problems of pedagogy and curricular design.

Over the past two decades we have created a vast and impressive infrastructure for assessment. It is strong and seemingly self-perpetuating, and we can be proud of it. Our challenge for the future will be to harness this infrastructure to partner actively with other reform initiatives on our campuses that are similarly situated—well-meaning, but largely isolated from one another—to help create an integrated approach to transforming teaching and learning. This kind of intentional linkage to a broader instructional reform agenda—as embodied in such documents as *Involvement in Learning*—is where assessment began some two decades ago. Reviving and fostering it, I believe, is the best way to escape “Perpetual Movement.”