Part One

History and Current Status of Assessment

In Chapter One, Peter Ewell, dean of the outcomes assessment movement in higher education, provides a comprehensive overview of this movement, from its beginnings in 1985 to the present. He examines the intellectual roots of assessment in such research traditions as those associated with the impact of college on student learning, program evaluation, and "scientific management." He suggests two directions for action designed to transform assessment from a movement into a culture and four kinds of scholarship that could move the field forward over the next several decades.

Marvin Peterson and Derek Vaughan have been involved in a three-phase study of the current status of assessment conducted as a component of the research agenda of the National Center for Postsecondary Improvement. The focus of their presentation in Chapter Two is the organizational and administrative patterns of support for assessment that promote the use of assessment information in making decisions aimed at improving institutional effectiveness. Although the national studies confirm the overall impression that extensive use of assessment as a vehicle for guiding improvement is under way in only a small number of institutions, those studies do suggest some characteristics of successful programs and three conceptual models of institutional support for assessment that can improve its use in the future.

Chapter One

An Emerging Scholarship: A Brief History of Assessment

Peter T. Ewell

This chapter offers a brief historical and analytical review of the assessment movement, from approximately 1985 to the present. It first examines some major events and forces influencing assessment's evolution as a "scholarship," including demands for curricular and pedagogical reform, shifting patterns of accountability, and changes in instructional delivery. It also examines significant scholarly themes and issues that have arisen in assessment's short history in such realms as epistemology, methodology, politics, and the use of information. The chapter concludes that assessment scholarship has become rich, robust, and strong. Whether it can or should continue as a distinct conversation outside the mainstream of higher education is more debatable.

Forerunners

The intellectual roots of assessment as a scholarship extend back well before its emergence as a recognizable movement. Some of its most visible forebears relate to undergraduate learning and the student experience in college. Others, such as program evaluation and "scientific management," helped direct its conscious orientation toward action and improvement. Methods and techniques drawn from these established traditions decisively influenced the language and methods of early assessment practitioners and continue to do so today.

Student Learning in College

This research tradition examines collegiate learning as a particular application of educational and developmental psychology. As such, its primary objective is discipline-based hypothesis testing and theory building, though its authors have often drawn implications for practice. Some of this work dates back to the 1930s and 1940s (for example, Learned and Wood, 1938), and much of it focused on single colleges enrolling eighteen- to twenty-one-year-old students in traditional residential environments. General maturation and attitudinal development were thus as much of interest as cognitive gain (Chickering, 1969). By the end of the 1960s there was a large enough body of work in this area for Feldman and Newcomb (1969) to synthesize its findings, which was updated some two decades later by Pascarella and Terenzini (1991). On the verge of assessment's emergence in the late 1970s, a trio of volumes was especially influential: Astin's *Four Critical Years* (1977) established the metaphor "value-added" and promoted the use of longitudinal studies to examine net effects, Bowen's *Investment in Learning* (1977) helped establish a public policy context for assessment by emphasizing the societal returns on investment associated with higher education, and Pace's *Measuring the Outcomes of College* (1979) emphasized the role of college environments and actual student behaviors. The contributions of this research tradition to assessment were both conceptual and methodological. Among the most prominent were basic taxonomies of outcomes, models of student growth and development, and tools for research like cognitive examinations, longitudinal and cross-sectional surveys, and quasi-experimental designs.

Retention and Student Behavior

Closely related to research on college student learning, a distinct literature on retention emerged in the late 1960s and 1970s and had some very specific impacts on assessment practice. First, it quickly organized itself around a powerful theoretical model—Tinto's notion of academic and social integration (1975), which proved equally useful in guiding applied research on student learn-

ing (for example, Terenzini, Pascarella, and Lorang, 1982). Second, the phenomenon of student attrition constituted an ideal proving ground for new methodologies involving longitudinal study designs, specially configured surveys, and multivariate analytical techniques, later adopted by many assessment practitioners. Third and perhaps decisively, retention scholarship was action research: though theoretically grounded and methodologically sophisticated, its object was always informed intervention (for example, Lenning, Beal, and Sauer, 1980). Together, these features yielded an excellent model of applied scholarship that, consciously or unconsciously, many assessment practitioners worked to emulate.

Evaluation and "Scientific Management"

The 1960s and 1970s also saw the rise of program evaluation as an action research tradition. Occasioned by the many large-scale federal programs launched at that time, program evaluation first relied largely on quantitative methods. It was also related to a wider movement toward "scientific management" that quickly found applications in higher education in the form of strategic planning, program review, and budgeting. The kind of "systems thinking" embedded in this tradition demanded explicit attention to student outcomes (for example, Enthoven, 1970) in order to provide a needed "output variable" for cost-benefit studies and investigations of social return on investment. This tradition also yielded one of the most extensive taxonomies of collegiate outcomes ever produced (Lenning, Lee, Micek, and Service, 1977) and stimulated a range of surveys designed to provide campuses with information about how students used and perceived their programs. Literature drawn from program evaluation further provided assessment with a ready-made set of models and vocabularies (for example, Light, Singer, and Willett, 1990). Somewhat later, program evaluation began to embrace more qualitative methods (for example, Guba and Lincoln, 1981). These more "authentic" approaches, which emphasized holistic examination of organizational situations and often employed open-ended interviewing and participant observation, also provided an early language for assessment for those skeptical of overly empirical methodologies.

Mastery Learning

The mastery and competency-based learning movement began in elementary and secondary education, but quickly found postsecondary applications in adult and professional education by the mid-1960s. Because mastery-based designs for learning are based entirely on agreed-upon outcomes, assessing and certifying individual student achievement was always paramount. A related development was the assessment of prior learning. Corporate assessment centers, meanwhile, were developing ways to examine and certify complex higher-order abilities by observing group and individual performance of authentic tasks (Thornton and Byham, 1982). Collectively, these traditions provided the conceptual foundation for "alternative" institutions like Empire State, Evergreen State, Regents College, Antioch College, and the School for New Learning at DePaul, as well as, and by far the most influential, Alverno College (Alverno College Faculty, 1979). They also yielded a cadre of early assessment practitioners, skilled in evaluating student portfolios and other authentic evidence of student attainment. Two contributions were especially important for the early assessment movement: first, mastery methods posed an effective alternative to the prominent (and politically popular) "testing and measurement" paradigm; second, they could boast a track record that proved that assessment in higher education was not just a popular theory; it could actually be done.

These four practice traditions and their associated literatures are quite different, and only a few educators in the early 1980s were reading them all. More significantly, their values and methodological traditions are frequently contradictory, revealing conceptual tensions that have fueled assessment discussions ever since. One is a clash of guiding metaphor between quantitative "scientific" investigation and qualitative "developmental" observation. Another addresses how assessment is positioned in the teaching-learning process: the "evaluation" and "measurement" traditions consciously divorce the process of investigating student attainment from the act of instruction in the name of objectivity; "mastery" tra-

ditions, in contrast, consider the two inseparable. A final distinction concerns the predominant object of assessment—whether its principal purpose is to examine *overall* program/institutional effectiveness or to certify what a *particular* student knows and can do. As any examination of early assessment citations will show, all four traditions helped shape language and practice in the early 1980s. What is surprising in retrospect is that such disparate scholarly traditions could be related at all and that they continue to inform such a lively scholarship.

Birth of a Movement

Although no one has officially dated the birth of the assessment movement in higher education, it is probably safe to propose that date as the First National Conference on Assessment in Higher Education, held in Columbia, South Carolina, in the fall of 1985. Cosponsored by the National Institute of Education (NIE) and the American Association for Higher Education (AAHE), the origins of this conference vividly illustrate the conflicting political and intellectual traditions that have been with the field ever since. The proximate stimulus for the conference was a report entitled *Involvement in Learning* (Study Group on the Conditions of Excellence in American Higher Education, 1984). Three main recommendations, strongly informed by research in the student learning tradition, formed its centerpiece. In brief, to promote higher levels of student achievement, it was recommended that high expectations be established for students, that students be involved in active learning environments, and that students be provided with prompt and useful feedback. But the report also observed that colleges and universities could "learn" from feedback on their own performance and that appropriate research tools were now available for them to do so.

This observation might have been overlooked were it not consistent with other voices. One set of voices came from within the academy and focused on curriculum reform, especially in general education. Symbolized by other prominent reports, like *Integrity in the College Curriculum* (American Association of Colleges, 1985) and *To Reclaim a Legacy* (Bennett, 1984), their central argument was that what was needed were coherent curricular experiences that

could be shaped by ongoing monitoring of student learning and development. From the outset in these discussions, the assessment of learning was presented as a form of scholarship. Faculties ought to be willing to engage in assessment as an integral part of their everyday work. A concomitant enlightened, but unexamined, assumption was that the tools of social science and educational measurement, deployed appropriately, could be adapted in all disciplines to further this process of ongoing inquiry and improvement.

A second set of voices arose simultaneously outside the academy, consisting largely of state-based calls for greater accountability. In part, these calls were a byproduct of the far more visible attention then being paid to K–12 education, symbolized by the U.S. Department of Education's report *A Nation at Risk* (1983). In part, they stemmed from a renewed activism by governors and legislatures, based on their growing recognition that postsecondary education was a powerful engine for economic and workforce development. Both themes were apparent in yet another national report—revealingly titled *Time for Results* (National Governors' Association, 1986). As it was being issued, Colorado and South Carolina adopted assessment mandates requiring public colleges and universities to examine learning outcomes and report what they found. (A few other states, such as Tennessee and Florida, for varying reasons, had been doing assessment for several years, using common standardized tests.) By 1987, when the first stocktaking of this growing policy trend occurred (Boyer, Ewell, Finney, and Mingle, 1987), about a dozen states had similar mandates. By 1989, this number had grown to more than half (Ewell, Finney, and Lenth, 1990).

Given this history, the motives of those attending the first national assessment conference were understandably mixed. Many were there under the banner of *Involvement in Learning* (Study Group on the Conditions of Excellence in American Higher Education, 1984), seeking reasonable and valid ways to gather information to improve curriculum and pedagogy. At least as many (and probably more) were there in response to a brand new mandate. Clear to all were the facts that they had few available tools, they had only a spotty literature of practice, and they had virtually no common intellectual foundation on which to build. Filling these yawning gaps in the period of 1985–1988 was a first and

urgent task for the scholarship of assessment. In beginning this task, practitioners faced three major challenges—of definitions, instruments, and implementation.

Definitions

One immediate problem was that the term *assessment* meant different things to different people. Initially, at least three meanings and their associated traditions of use had therefore to be sorted out. The most established definition had its roots in the mastery-learning tradition, where assessment referred to the processes used to determine an individual's mastery of complex abilities, generally through observed performance (for example, Alverno College Faculty, 1979). Adherents of this tradition emphasized development over time and continuous feedback on individual performance, symbolized by the etymological roots of the word *assessment* in the Latin *ad* + *sedere,* "to sit beside" (Loacker, Cromwell, and O'Brien, 1986). A far different meaning emerged from K–12 practice, where the term described large-scale testing programs like the federally funded National Assessment of Educational Progress (NAEP) and a growing array of state-based K–12 examination programs. The primary objective of such "large-scale assessment" was not to examine individual learning but rather to benchmark school and district performance in the name of accountability. Its central tools were standardized examinations founded on well-established psychometric principles, designed to produce summary performance statistics quickly and efficiently. Yet a third tradition of use defined assessment as a special kind of program evaluation, whose purpose was to gather evidence to improve curricula and pedagogy. Like large-scale assessment, this tradition focused on determining aggregate, not individual, performance, employing a range of methods, including examinations, portfolios and student work samples, surveys of student and alumni experiences, and direct observations of student and faculty behaviors. An emphasis on improvement, moreover, meant that assessment was as much about *using* the resulting information as it was about psychometric standards.

All three definitions raised explicitly the dichotomy of purpose apparent from the outset: accountability versus improvement. Other differences addressed methods and units of analysis—essentially

whether quantitative or qualitative methods would predominate and whether attention would be directed largely toward aggregate or individual performance. Clarifying such distinctions in the form of taxonomies helped sharpen initial discussions about the meaning of assessment (Terenzini, 1989). They also helped further a terminological consensus that was centered on the use of multiple methods for program improvement (American Association for Higher Education, 1992).

Instruments

A second challenge faced by early assessment practitioners was the task of quickly identifying credible and useful ways of gathering evidence of student learning. Virtually all the available instruments were designed for something else. Ranging from admissions tests like the ACT Assessment and the Graduate Record Examinations, through professional registry and licensure examinations, to examinations designed to award equivalent credit, none of the available testing alternatives was really appropriate for program evaluation. Their content only approximated the domain of any given institution's curriculum, and the results they produced usually provided insufficient detail to support improvement. But this did not prevent large numbers of institutions—especially those facing state mandates—from deploying them. One exception was the ACT College Outcomes Measures Project (COMP) examination (Forrest and Steele, 1978). In many ways a harbinger, this examination was designed specifically to evaluate general education outcomes and support group-level inferences about student learning. It also constructed general education outcomes in novel ways, emphasizing the application of knowledge in real-world situations and (in its long form) requiring authentic demonstrations of performance.

In the period of 1986–1989, the major testing organizations quickly filled the instrument gap with a range of new purpose-built group-level examinations aimed at program evaluation—all based on existing prototypes. Among the most prominent were the ACT Collegiate Assessment of Academic Proficiency (CAAP), the Educational Testing Service (ETS) Academic Profile, and a range of ETS Major Field Achievement Tests (MFAT). Student surveys provided another readily available set of data-gathering tools, espe-

cially when they contained items on self-reported gain. While many institutions designed and administered their own surveys, published instruments were readily available, including the Cooperative Institutional Research Program (CIRP) Freshman and follow-up surveys, the College Student Experiences Questionnaire (CSEQ), and a range of questionnaires produced by organizations like ACT and the National Center for Higher Education Management Systems (NCHEMS).

The principal appeal of off-the-shelf tests and surveys in this period was their ready availability—a property enhanced when the first comprehensive catalogs of available instruments appeared (Smith, Bradley, and Draper, 1994). Faced with a mandate demanding immediate results, most institutions felt they had little choice but to use such instruments, at least in the short term. But there were also growing doubts about the wisdom of this approach (Heffernan, Hutchings, and Marchese, 1988), stimulating work on more authentic, faculty-made assessment approaches in the coming years.

Implementation

A third challenge faced by early assessment practitioners was the lack of institutional experience about how to carry out such an initiative. One question here concerned cost, and as a result, some of the first "how to" publications addressed financial issues (Ewell and Jones, 1986). Others considered the organizational questions involved in establishing an assessment program (Ewell, 1988). But absent any real exemplars, the guidance provided by such publications was at best rudimentary. Enormous early reliance was therefore placed on the lessons that could be learned from the few documented cases available. Three such early adopters had considerable influence. The first was Alverno, whose "abilities-based" curriculum, designed around performance assessments of every student, was both inspiring and daunting (Alverno College Faculty, 1979). A second early adopter was Northeast Missouri (now Truman) State University, which since 1973 had employed a range of nationally normed examinations to help establish the "integrity" of its degrees (McClain, 1984). A third was the University of Tennessee, Knoxville, which, under the stimulus of Tennessee's performance funding scheme, became the first major

public university to develop a comprehensive multimethod system of program assessment (Banta, 1985). These three cases were very different and provided a wide range of potential models. They were also unusually well documented, yielding some of the first concrete examples of assessment scholarship.

In the late 1980s a second wave of documented cases emerged, including (among others) James Madison University, Kean College, Kings College, Ball State University, Miami-Dade Community College, and Sinclair Community College—many of which were responding to new state mandates. To a field hungry for concrete information, these examples were extremely welcome. More subtly, they helped define a "standard" approach to implementing a campus-level program, which was widely imitated.

This founding period thus generated some enduring lines of assessment scholarship. One line of work addressed concept development and building a coherent language. The purpose here was largely to stake out the territory—though much of this early literature was frankly hortatory, intended to persuade institutions to get started. A second line of work concerned tools and techniques, and though all "forerunner" literatures were referenced here, strong reservations about standardized testing quickly emerged and persisted. A third strand comprised case studies of implementation, supplemented by a growing body of work addressing practical matters like organizational structures and faculty involvement. Finally, accountability remained a distinct topic for comment and investigation, looking primarily at state policy but shifting later to accreditation.

Into the Mainstream

By 1990, predictions that "assessment would quickly go away" seemed illusory. Most states had assessment mandates, though these varied in both substance and the vigor with which they were enforced. Accrediting bodies, meanwhile, had grown in influence, in many cases replacing states as the primary external stimulus for institutional interest in assessment (Ewell, 1993). Reflecting this

shift, more and more private institutions established assessment programs. These external stimuli were largely responsible for a steady upward trend in the number of institutions reporting involvement with assessment. For example, in 1987 some 55 percent of institutions claimed that they had established an assessment program on the American Council of Education's (ACE) annual *Campus Trends* survey. By 1993, this proportion had risen to 98 percent (though the survey also suggested that most such efforts were only just getting started). Clearly, at least for administrators, assessment was now mainstream. But entering the mainstream meant more than just widespread reported use. It also implied consolidation of assessment's position as a distinct and recognizable scholarship of practice.

An Emerging Modal Type

As institutions scrambled to "implement assessment," it was probably inevitable that they would evolve similar approaches. And despite repeated admonitions to ground assessment in each institution's distinctive mission and student clientele, they approached the task of implementation in very similar ways. As a first step, most formed committees to plan and oversee the work. Following widespread recommendations about the importance of faculty involvement, most comprised faculty drawn from multiple disciplines. But partly because the press to implement was so great, assessment committees rarely became a permanent feature of governance or of academic administration.

The clear first task of these committees, moreover, was to develop an assessment plan. Often, such a product was explicitly required by an accreditor or state authority. Equally often, it was recommended by a consultant or by the burgeoning "how to" literature of practice (for example, Nichols, 1989). The resulting plans thus often had a somewhat formulaic quality. Most included an initial statement of principles, stated learning goals for general education and for each constituent discipline, a charge to departments to find or develop a suitable assessment method (frequently accompanied by a list of methods to be considered), and a schedule for data collection and reporting. Implementing such plans, in turn, often involved the use of specially funded "pilot" efforts by

volunteer departments. Keeping track of implementation and reporting, moreover, often demanded the use of a tabular or matrix format (Banta, 1996), and this, too, became a widespread feature of the "standard" approach. Methods, meanwhile, were healthily varied, including available standardized examinations, faculty-made tests, surveys and focus groups, and (increasingly, as the decade progressed) portfolios and work samples.

A Literature of Practice

In assessment's early days, the products of its scholarship comprised a fugitive literature of working papers, loosely organized readings in *New Directions* sourcebooks, and conference presentations. But by the early 1990s, the foundations of a recognizable published literature could be discerned. Some of these works were by established scholars, who summarized findings and provided methodological advice (Astin, 1991; Pace, 1990). Others tried to document assessment approaches in terms that practitioner audiences could readily understand (Erwin, 1991; Ewell, 1991a). Still others continued the process of documenting institutional cases—of which there were now many—in standard or summary form (Banta and Associates, 1993).

The establishment of the movement's own journal, *Assessment Update,* in 1989, was also an important milestone in this period—providing relevant commentary on methods, emerging policies, and institutional practices. As its editorial board envisioned, its contents were short, practical, and topical, providing the field with a single place to turn for ideas and examples. *Assessment Update*'s existence also provided an important alternative to established educational research journals for faculty-practitioners who wanted to publish. This supplemented the already-established role of *Change* magazine, which provided an early venue for assessment authors and continued to print assessment-related essays regularly (DeZure, 2000). Through its Assessment Forum, moreover, AAHE issued a range of publications, building first upon conference presentations and continuing in a set of resource guides (American Association for Higher Education, 1997). In strong contrast to fifteen years previously, assessment practitioners in 2000 thus had a significant body of literature to guide their efforts, which included

systematic guides to method and implementation (for example, Palomba and Banta, 1999), well-documented examples of campus practice (for example, Banta, Lund, Black, and Oblander, 1996), and comprehensive treatises integrating assessment with the broader transformation of teaching and learning (for example, Mentkowski and Associates, 2000).

Scholarly Gatherings and Support

Initiated on a regular annual cycle in 1987, the AAHE Assessment Forum was by 1989 *the* conference for practitioners, providing a regular gathering place for scholarly presentation and exchange. Sessions developed for the Forum required formal documentation and often ended up as publications. The Forum also maintained professional networks, promoted idea sharing, and provided needed moral support and encouragement. The latter was especially important in assessment's early years because there were few practitioners and they were isolated on individual campuses. Although the Forum remained the field's premier conference, other gatherings quickly emerged. Some, like the Assessment Institute in Indianapolis (which actually began at the University of Tennessee, Knoxville), concentrated largely on orienting new practitioners. Others arose at the state level, including (among others) the South Carolina Higher Education Assessment Network (SCHEA), the Washington Assessment Group (WAG), and the Virginia Assessment Group (VAG), and often were directly supported by state higher education agencies. Some of these state-level groups published regular newsletters updating members on state policy initiatives and allowing campuses to showcase their programs. Funding support for assessment scholarship also became more accessible, primarily through the federal Fund for the Improvement of Postsecondary Education (Cook, 1989). In addition to directly supporting assessment activities, FIPSE's need for formal reports and evaluations helped stimulate the field's growing inventory of published work.

A "Semi-Profession"

Although assessment remained largely a part-time activity, entering the mainstream also meant a rise in the number of permanent

positions with assessment as a principal assignment. Position titles like "assessment coordinator," with formal job descriptions, are now commonplace, usually located in academic affairs offices or merged with institutional research. The creation of such positions was in large measure a result of external pressure to put recognizable campus programs in place. Certainly, such roles helped build badly needed local capacity and infrastructure. But in many cases they also created real tensions about the ownership and benefits of the assessment process.

Early conversations, meanwhile, considered the advisability of creating a national professional organization for assessment similar to the Council for Adult and Experiential Learning (CAEL). A strong consensus emerged to maintain assessment as an "amateur" activity—undertaken by faculty themselves for the purpose of improving their own practice. Avoiding excessive professionalization was important because it promoted later linkages with the scholarship of teaching. But large and growing numbers of individuals on college and university campuses, often without conscious choice, have nevertheless adopted careers identified primarily with assessment as a distinguishable field.

For assessment as a whole, one clear result of entering the mainstream is an established community of practice that in some ways resembles an academic discipline. Among its earmarks are an identifiable and growing body of scholarship, a well-recognized conference circuit, and a number of "sub-disciplines," each with its own literature and leading personalities. Certainly, this is a significant achievement—far beyond what numerous early observers expected. But these very attributes have also decisively shaped, and in some ways limited, assessment's impact on instruction and campus culture. Most campus assessment activities, for example, continue to be implemented as *additions* to the curriculum, designed for purposes of program evaluation rather than being integral to teaching and learning. The fact that implementation so often centers on "doing assessment" rather than on improving practice through clear linkages to budget and pedagogy, moreover, can easily isolate the process from the everyday life of both faculty and administrators. Those doing assessment have evolved a remarkably

varied and sophisticated set of tools and approaches and an effective semiprofessional infrastructure to support what they do. But few faculty as yet practice assessment as a part of their everyday work. Although firmly established in the mainstream by the year 2000, assessment as a movement is still striving for the cultural shift its original proponents had hoped for.

Episodes and Debates

Throughout its brief history, assessment has addressed a varied set of intellectual issues that have actively stimulated debate. Meanwhile, the movement went through several telling episodes that forced reaction and rethinking. Each episode prompted deeper understanding, though none has been entirely resolved. As a result, past events continue to influence the course of this evolving scholarship.

The "Ineffability" Debate

Perhaps the most basic debate that arises as faculty face assessment is the extent to which educational outcomes can be specified and measured at all. Indeed, a frequent early counterargument was that *any* attempt to look at outcomes directly was both demeaning and doomed to failure. Related critiques noted that assessment's principal vocabulary appeared confined to education and the social sciences—not always the most respected disciplines on any college campus. More pointedly, both the rhetoric and the implied methods advanced by the assessment movement have frequently been characterized as "positivist" and excessively mechanistic. Dissecting this classic complex of faculty reservations about assessment reveals some quite different underlying issues. Some are legitimately methodological, including appropriate reservations about the ability of off-the-shelf instruments and forced-choice methods to fully reflect collegiate learning, or fears about "teaching to the test." Some are profoundly philosophical, based on a recognition that deep learning is always holistic, reflective, and socially constructed. Still others are predominantly political, derived from faculty fears about loss of autonomy and creeping management control, as well as concerns about external intrusion into the curriculum.

What makes things complicated is that all three issues are often bound up in a single sense of discomfort (for example, Peters, 1994). But the resulting debate about "ineffability" has proven helpful in deepening assessment scholarship. At one level, it forced practitioners to sharpen the philosophical grounding of the movement—rooting it in the tenets of scholarship and the process of teaching and learning. It also reemphasized that the evidence used by assessment must always rest upon a peer-based community of judgment (Mentkowski and others, 1991; American Association for Higher Education, 1992). Finally, the debate forced explicit recognition of the fact that evidence is consistently constrained by the context in which it is generated (Mentkowski and Rogers, 1988) and by the uses to which it is put (Messick, 1988). Epistemological issues of this kind thus remain at the heart of the movement and remain healthily and vigorously contested (Ewell, 1989; Harris and Sansom, 2001). But protests based solely on principle or politics have steadily diminished.

The "Value-Added" Debate

The question of whether assessment's primary focus should be placed on documenting absolute levels of student attainment or on institutional contributions to developing student abilities arose early (Ewell, 1984). Reasons for centering attention on "talent development" were compelling. First, this approach was normatively appealing and had the admirable property of leveling the playing field among different kinds of institutions. Assessing institutional quality in this way thus made more sense than using traditional markers like resources and reputation (Astin, 1985). Ascertaining "net effects" also made good sense from a research point of view, recognizing that incoming student ability is the largest predictor of any outcome (Pascarella, 1987; Pascarella and Terenzini, 1991). Finally, some institutions were already practicing "value-added" approaches and finding them useful in demonstrating effectiveness (McClain and Krueger, 1985).

But the classic approach to assessing learning gain—testing students on entry and then retesting them at exit—posed perplexing conceptual issues and formidable methodological problems. Con-

ceptually, it was argued, a pretest was often simply silly because students had not yet been exposed to the subject on which they were being tested (Warren, 1984). The term *value-added,* moreover, suggested a mechanistic view of education, in which students were viewed as "products" and learning merely additive. Actually determining growth, meanwhile, entailed multiplicative sources of measurement error and sometimes led to real misinterpretations of underlying phenomena (Hanson, 1988; Baird, 1988; Banta and others, 1987). Although active discussion of this topic diminished in the 1990s, it helped propel assessment toward a useful synthesis. Most important, these discussions helped forge a growing consensus that paths of student development should not be seen as linear and additive but rather as organic and transformational. A methodological entailment of this growing consensus was longitudinal designs for assessment, capable of capturing large numbers of variables about both outcomes and experiences. Such longitudinal studies required an analytical model based on multivariate statistical control instead of simple "test-retest" approaches, and could be further enhanced by the use of qualitative methods like periodic interviews and focus groups. Finally, all agreed that for policy purposes, information about *both* levels of attainment and institutional contributions was needed.

The TQM Episode

In the early 1990s higher education institutions began experimenting with Total Quality Management (TQM), a set of ideas and techniques borrowed directly from business, to help improve their administrative operations (Seymour, 1991). Linking such notions with assessment was appealing because the two movements shared many attributes. Both began with a systemic approach to change and, indeed, viewed change itself as imperative. Both emphasized the need to listen carefully to those whom the system was trying to serve, although the notion of students as "customers" immediately grated. Finally, both held that concrete information about performance was a critical part of a continuous cycle of planning and improvement. Recognizing such parallels, AAHE incorporated a track on TQM—quickly relabeled Continuous Quality Improvement

(CQI)—into its Assessment Forum and issued a number of publications linking assessment and CQI (for example, American Association for Higher Education, 1994a).

But explicit attempts to fuse assessment and Total Quality were not successful, and after its initial flurry of activity, Total Quality has not fared well on campuses. Partly, this was a matter of language. Whereas assessment could ultimately adopt the discourse of scholarship, Total Quality never shed its corporate flavor—especially in the eyes of skeptical faculty. Partly, it was because the quality movement in business and industry itself had peaked. Yet much was synthesized by assessment—perhaps unconsciously—from this encounter. It reinforced "systems consciousness" and cemented the need to collect information about both outcomes and processes. Sometimes bitter "customer" discussions helped underline the need to listen carefully to student voices and to shift assessment's perspective from faculty "teaching" to student learning. Total Quality thus proved useful to assessment largely as a metaphor. At the same time, it taught object lessons about the risks of both alien language and the appearance of fad.

The National Assessment Episode

In 1990, the National Education Goals Panel established the nation's first objectives for collegiate learning. More specifically, it called for the development of valid and reliable assessments to track progress in critical thinking, communication, and problem solving (National Education Goals Panel, 1991). This action mirrored simultaneous and growing state interest in collegiate assessment, as a majority of the states had adopted assessment mandates for public colleges and universities by the mid-1990s. The Goals also signaled the beginning of a significant, though short-lived, period of aggressiveness by the U.S. Department of Education within the realm of postsecondary accountability, marked by such initiatives as the Student Right-to-Know Act and the State Postsecondary Review Entities (SPREs). Like the latter, the proposed national assessment never happened, though a major design effort to create it helped stimulate useful thinking about how a large-scale, authentic assessment of collegiate learning might actually be deployed (National Center for Education Statistics, 1994). Similar

calls for a "NAEP for College" have periodically arisen, stimulated by both accountability demands and international comparisons (National Center for Public Policy and Higher Education, 2000).

To be sure, such episodes have had but little impact on the day-to-day practice of assessment on most college campuses. But together with their state-level counterparts, they provide a constant reminder that accountability was part of assessment's birthright and is intimately entwined with its future. Continuing scholarship aimed at developing appropriate and timely responses to periodic accountability demands as they arise will therefore always be needed.

Each of these issues illustrates how assessment discourse has grown in sophistication and has built a particular set of shared understandings. All remain centers of active debate. But few are now posed in the black-and-white terms in which they first arose, and each has helped stimulate improvements in methods and approach.

Into the Future

Social and educational movements, whatever their object, have one of two typical fates. Unsuccessful movements vanish after only a few years, with little left behind. Successful ones disappear equally as "movements" because their core values become part of the dominant culture and their practices are fully institutionalized. So far, the assessment movement has experienced neither. On the one hand, levels of activity are unprecedented. The vast majority of institutions continue to report engagement with assessment, conference attendance is burgeoning, publications abound, and a growing body of practitioners see assessment as their primary professional practice. On the other hand, at most institutions—and above all, for most individual faculty—assessment has not become a "culture of use" (López, 1997). The resulting paradox raises two questions, both highly relevant to the movement's future. First, what is it that has sustained assessment for so long, and what will continue to do so? Second, what has prevented assessment from fulfilling its original promise, and how might it ultimately achieve

these ends? Answers to both these questions, admittedly, are speculative and uncertain. But, as the answers eventually play out, they will decisively affect the scholarship of assessment.

Why Didn't Assessment Go Away?

In assessment's first decade, the question *When will it go away?* was frequently posed. This was largely because the movement was diagnosed by many as a typical "management fad"—like Total Quality or Management by Objectives (MBO)—that would quickly run its course (Birnbaum, 2000). Yet assessment has shown remarkable staying power and has undoubtedly attained a measure of permanence, at least in the form of a visible infrastructure. Several factors appear responsible for this phenomenon. Probably the most important is that external stakeholders will not let the matter drop. State interest is now stronger than ever, fueled by demand-driven needs to improve "learning productivity" and by burgeoning state efforts to implement standards-based reform in K–12 education (Ewell, 1997b). Accreditation agencies, meanwhile, have grown increasingly vigorous in their demands that institutions examine learning outcomes, though they are also allowing institutions more flexibility in how they proceed (Eaton, 2001). Market forces and the media are not only more powerful, they are also far more performance-conscious and data-hungry than they were two decades ago. Assessment thus has become an unavoidable condition of doing business: institutions can no more abandon assessment than they can do without a development office.

The last twenty years have also seen a revolution in undergraduate instruction. In part, this results from technology, and in part, it reflects the impact of multiple other movements, including writing across the curriculum, learning communities, problem-based learning, and service learning. Together, these forces are fundamentally altering the shape and content of undergraduate study. Such changes are sustaining assessment in at least two ways. Most immediately, new instructional approaches are forced to demonstrate their relative effectiveness precisely because they are new. Assessment activities therefore are frequently undertaken as an integral part of their implementation. More subtly, the very nature of these new approaches shifts the focus of attention from

teaching to learning. In some cases, direct determination of mastery is integral to curricular design (O'Banion, 1997). In others, common rubrics for judging performance are required to ensure coherence in the absence of a more visible curricular structure (Walvoord and Anderson, 1998). Assessment has thus been sustained in part because it has become a necessary condition for undertaking meaningful undergraduate reform—just as the authors of *Involvement in Learning* (Study Group on the Conditions of Excellence in American Higher Education, 1984) foresaw.

Why Broad but Not Deep?

As important as assessment's longevity, though, is the fact that it has survived in a peculiar form. Most campuses are indeed "doing something" in assessment. But the kinds of fundamental transformations in instruction that might have resulted from examining systemic evidence of student learning have mostly not happened. Instead, for the majority of institutions, assessment remains an add-on, done principally at the behest of the administration and sustained as a superstructure outside the traditional array of academic activities and rewards. Reasons for this widespread condition, ironically, mirror those that have sustained assessment for almost two decades. First, widespread and visible external demands generally set the tone for initial engagement. Most campuses still do assessment because somebody tells them to. Regardless of how the telling is done (and external bodies have been more sensitive and flexible than is usually acknowledged), responses risk being both reactive and mechanistic.

Moreover, like other efforts to accomplish meaningful undergraduate reform, assessment must usually be implemented across the grain of deeply embedded organizational structures. Rewards for engaging in it remain scant for both institutions and individuals. So, like similar activities not rooted in disciplines or departments such as first-year experience programs or general education, assessment is frequently sustained as a *separate* activity, ensconced in an "office" and nurtured through special-purpose funding. Similarly, as Peterson and Vaughan report in Chapter Two, assessment results are rarely central to institutional planning and decision making, even when undertaken outside

the glare of public scrutiny. Partly, this is because of continuing faculty fears about negative consequences. Ironically, it stems equally from faculty expectations of no consequences at all—that considerable effort will be expended gathering information that will never be used. Much of the appeal of the kinds of activities that *have* been adopted on a widespread basis, like classroom assessment, is that the benefits of feedback are both immediate and apparent (Angelo and Cross, 1993). At the institutional and program levels, the benefits of assessment have been far less immediately visible.

As this last observation suggests, two fundamental changes will be needed to transform assessment from a movement into a culture. One is at the level of teaching and learning and requires shifting assessment's conceptual paradigm from an evaluative stance that emphasizes checking up on results to an emphasis on assuming active and collective responsibility for fostering student attainment. Forces that might aid this conceptual transformation include the growing salience of ability-based credentials, which are fast becoming a way of life in many occupations and professions (Adelman, 2000). Multi-institutional attendance patterns are meanwhile fueling demands to reposition articulation and transfer from course-based "seat time" to performance-based attainment. Perhaps most important, reform efforts like writing across the curriculum and problem-based learning, together with technology, are forcing faculty to think far more concretely and collectively about learning outcomes and how to certify them.

A second needed transformation is at the level of academic administration and requires evolving a largely top-down, management-oriented use of information in planning and decision making toward a culture that more fully embodies the principles of a learning organization. Forces that might help this transformation are far less easy to identify but include growing competition from nonuniversity providers and insistent demands to create modes of instruction that are both efficient and effective.

Such developments, if they occur, will influence the scholarship of assessment profoundly. In its literature, there will likely be growing sophistication in discussions of methodology, capitalizing on emerging knowledge about how to forge consensual judgments about authentic performance (for example, Walvoord and Anderson,

1998; Mentkowski and Associates, 2000) and about using technology to deliver assessments based on complex, interactive problems. There will be new demands for work on organizational transformation, in which assessment is addressed but is fused with other systemic changes aimed at changing the environment for teaching and learning (for example, Gardiner, 1994; O'Banion, 1997; Harvey and Knight, 1996). At a different level, assessment will gradually become an integral part of each faculty member's reflective practice, documented through the scholarship of teaching (Shulman, 1993; Hutchings, 1996). And faculty will increasingly collaborate in this work, reflecting their growing assumption of collective responsibility for learning.

Such developments are consistent with the tradition of robust, participatory, and practice-oriented scholarship already established by the assessment movement. If they emerge, they will constitute significant contributions to both theory and practice. More important, because assessment's principal tenets will at last be embodied in the work of higher education on a day-to-day basis, its life as a "movement" may finally be over.