ENHANCING ASSESSMENT IN HIGHER EDUCATION

Putting Psychometrics to Work

Edited by

Tammie Cumming and

M. David Miller

Foreword by Michael J. Kolen

Copublished in association with

AIR ASSOCIATION FOR INSTITUTIONAL RESEARCH
Data and Decisions for Higher Education

A|A|C|U Association of American Colleges and Universities

Stylus

STERLING, VIRGINIA

Sty/us

Bulk Purchases

Quantity discounts are available for use in workshops and for
staff development.
Call 1-800-232-0223

HISTORY AND CONCEPTUAL BASIS OF ASSESSMENT IN HIGHER EDUCATION

Peter T. Ewell, National Center for Higher Education Management Systems

Tammie Cumming, City University of New York–New York City College of Technology

In introducing the volume, this chapter has two main purposes. Its first purpose is to offer a brief historical and analytical review of major events and forces influencing the evolution of assessment from 1985 to date including demands for curricular and pedagogical reform, shifting patterns of accountability, and changes in instructional delivery. At the same time, it examines significant methodological and conceptual issues that have arisen in assessment's short history in such realms as epistemology, assessment design, psychometric properties, politics, and the use of assessment information. Its second purpose is to briefly review key concepts, terms, and approaches that characterize the current practice of assessment. Some of these, to be sure, have historical roots and have evolved a good deal since they were introduced decades ago. But all are robust components of student learning outcomes assessment as practiced today.

The Evolution of Assessment in Higher Education

Assessment of student learning has evolved throughout the years. In this chapter, we explore the historical basis for modern assessment techniques and usage in higher education and accreditation.

Some Intellectual Forbearers

The intellectual roots of assessment as a distinct collection of concepts and methods for gathering evidence extend well before its emergence as a recognizable movement. Some of its most visible forbearers are about undergraduate learning and the student experience in college. Others helped ground 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. In parallel, each of these traditions left its distinctive methodological footprint affecting how evidence of student learning is gathered and how it is handled in analysis. The historical discussion in this chapter updates Ewell (2002) and develops additional methodological implications.

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 (e.g., Learned & Wood, 1938), and much of it focused on single colleges enrolling 18- to 21-year-old students in traditional residential environments. General maturation and attitudinal development was thus as much an area of interest as cognitive gain (e.g., 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, updated some two decades later by Pascarella and Terenzini in 1991 and again in 2005. On the verge of assessment's emergence in the late 1970s, a trio of volumes was especially influential: Astin's (1977) *Four Critical Years* established the idea of "value added" and promoted the use of longitudinal studies to examine net effects; Bowen's (1977) *Investment in Learning* helped establish a public policy context for assessment by emphasizing the societal returns on investment associated with higher education; and Pace's (1979) *Measuring the Outcomes of College* emphasized the role of college environments and student behaviors, providing the conceptual roots of the National Survey of Student Engagement (NSSE). The contributions of this research tradition to assessment were both theoretical 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 the 1970s and had very specific

impacts on assessment practice. First, it quickly organized itself around a powerful theoretical model—Tinto's (1975) notion of academic and social integration—which proved equally useful in guiding applied research on student learning. 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 that were later adopted by many assessment practitioners. Third and perhaps decisively, retention scholarship was *action* research: Although theoretically grounded and methodologically sophisticated, its object was always informed intervention (e.g., Lenning, Beal, & Sauer, 1980). Together, these features yielded an excellent model of applied research 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 almost entirely 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 (e.g., 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, & Service, 1977). Literature drawn from program evaluation further provided assessment with a ready-made set of models and vocabularies (e.g., Light, Singer, & Willett, 1990). Somewhat later, program evaluation began to embrace more qualitative methods (e.g., Guba & 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 entirely based on agreed-upon outcomes, assessing and certifying individual student achievement was always paramount. A related development was the assessment of prior learning (Whitaker, 1989). Corporate assessment centers, meanwhile, were developing ways to examine

and certify complex higher order abilities by observing group and individual performances on authentic tasks (Thornton & Byham, 1982). Collectively, these traditions provided the conceptual foundation for "alternative" institutions like Empire State; Evergreen State; Regents College; Antioch College; the School for New Learning at DePaul; 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 that had its roots in K–12 accountability programs. 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 in the early 1980s were reading all of them. 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. Others address how assessment is positioned in the teaching–learning process: The "evaluation" and "measurement" traditions, for example, consciously divorce the process of investigating student attainment from the act of instruction in the name of objectivity; "mastery" traditions, 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 and methodological traditions would come together after more than 30 years as a recognized and reasonably well-integrated amalgam of theory and practice.

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 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 called *Involvement in Learning* (NIE, 1984). Three main recommendations formed its centerpiece, strongly informed by research in the student learning tradition. In brief, they were that higher levels of student achievement could be promoted by establishing high expectations for students, by involving students in active learning environments, and by providing students with prompt and useful feedback. But the report also observed that colleges and universities as institutions could learn from feedback on their own performances 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. The first 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* (Association of American Colleges & Universities, 1985) and *To Reclaim a Legacy* (Bennett, 1984), this group's central argument was the need for coherent curricular experiences that could best 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 by all disciplines to further this process of ongoing inquiry and improvement.

The 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 (1983) report *A Nation at Risk*. In part, it 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, states like Colorado and South Carolina adopted assessment mandates requiring public colleges and universities to examine learning outcomes and report what they found. By 1987, when the first stocktaking of this growing policy trend occurred (Boyer, Ewell, Finney, & Mingle, 1987), about a dozen states had similar mandates. By 1989, this number had grown to more than half the states (Ewell, Finney, & 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* (NIE, 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, only a spotty literature of practice, and virtually no common intellectual foundation on which to build. Filling these yawning gaps between 1985 and 1988 was a first and urgent task. In beginning this task, practitioners faced the following challenges.

Definitions

The first immediate challenge was that the term *assessment* meant different things to different people. Initially, at least three meanings and their associated traditions of use had to be sorted out. The first and most established 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 (e.g., 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, & O'Brien, 1986). A second, 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 object 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 rather than individual performance and employed 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 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). The taxonomies also helped further a terminological consensus centered on the use of multiple methods for program improvement. The *Standards for Educational and*

Psychological Testing (AERA, APA, & NCME, 2014; subsequently referred to as the *Standards*), the gold standard in educational measurement, notes that assessment is sometimes distinct from testing and refers to a broader process that integrates assessment information with information from other sources or it can consist of a single procedure in which a sample of examinee behavior is obtained and scored in a standardized manner.

Instruments

The second challenge faced by early assessment practitioners was to quickly identify credible and useful ways to gather evidence of student learning. Virtually all available instruments were designed to do something else. Ranging from admissions tests like the ACT college readiness 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 were 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.

In the period from 1986 until 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) Proficiency Profile (formerly known as the Academic Profile), and a range of ETS Major Field Achievement Tests (MFAT). Student surveys provided another readily available set of data-gathering tools, especially when they contained items on self-reported gain. Although many institutions designed and administered their own surveys, published instruments were readily available including the CIRP Freshman Survey and follow-up surveys, the College Student Experiences Questionnaire (CSEQ), and a range of questionnaires produced by organizations like ACT and National Center for Higher Education Management Systems.

The principal appeal of off-the-shelf tests and surveys was their ready availability—a property enhanced when the first comprehensive catalogues of available instruments appeared (Smith, Bradley, & 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, & Marchese, 1988), stimulating work on more authentic, faculty-made assessment approaches in the coming years.

Implementation

The third challenge faced by early assessment practitioners was lack of institution-level experience about how to carry out such an initiative. One concern here was cost and, as a result, some of the first how-to publications addressed financial issues (Ewell & Jones, 1986). Others considered the organizational questions involved in establishing an assessment program (Ewell, 1988). But absent 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 College, whose abilities-based curriculum designed around performance assessments of every student was both inspiring and daunting (Alverno College Faculty, 1979). The second 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). The 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.

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 for assessment's later development. One addressed concept development and building a coherent language. The purpose here was largely to stake out the territory—although much of this early literature was, frankly, hortatory, intended to persuade institutions to get started. A second line of work concerned tools and techniques. 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 toward accreditation.

Into the Mainstream

By 1990, predictions that assessment would quickly go away seemed illusory. Most states had assessment mandates, although 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% of institutions claimed they had established an assessment program on ACE's annual *Campus Trends* survey. By 1993, this proportion had risen to 98% (although the survey also suggested that most such efforts were only just getting started). Clearly, at least for administrators, assessment was now in the 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 practice.

An Emerging Modal Type
As institutions scrambled to "implement assessment," it was probably inevitable that they evolved 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 an initial 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 pressure to implement was so great, assessment committees rarely became a permanent feature of governance or of academic administration.

The clear starting 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 (e.g., Nichols, 1989). The resulting plans thus often had a somewhat formulaic quality. Most, for example, included (a) an initial statement of principles, (b) a list of stated learning goals for general education and for each academic program, (c) a charge to departments to find or develop a suitable assessment method (frequently accompanied by a list of methods to be considered), and (d) 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 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, practices and experiences were recorded in 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, 1991). Still others continued the process of documenting institutional cases—of which there were now many—in standard or summary form (Banta & Associates, 1993).

The establishment of the movement's own publication, *Assessment Update*, in 1989 was also an important milestone—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 regularly print assessment-related essays (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 (AAHE, 1997). In strong contrast to the 15 previous years, assessment practitioners in 2000 had a significant body of literature to guide their efforts that included systematic guides to method and implementation (e.g., Palomba & Banta, 1999), well-documented examples of campus practice (Banta, Lund, Black, & Oblander, 1996), and comprehensive treatises integrating assessment with the broader transformation of teaching and learning (e.g., Mentkowski & 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 sharing of ideas, 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. Other conferences arose at the state level including (among others) the South Carolina Higher Education Assessment (SCHEA) Network, the Washington Assessment Group (WAG), and the Virginia Assessment Group (VAG)—often

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. When the AAHE Assessment Forum ceased with the demise of its parent organization, its place was soon taken by the Assessment Institute in Indianapolis, with attendance figures topping 1,500. This gathering was joined by other important conferences at North Carolina State University; Texas A&M; and California State University, Long Beach by 2010.

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 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 so that accreditors could notice them. Certainly such roles helped build badly needed local capacity and infrastructure.

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 this evolution today 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 subdisciplines, each with its own literature and leading personalities. Those doing assessment, moreover, have evolved a remarkably varied and sophisticated set of tools and approaches and an effective semiprofessional infrastructure to support what they do. These are significant achievements—far beyond what numerous early observers expected.

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 Management (TQM)

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 remains strong, fueled by demand-driven needs to improve "learning productivity" and by burgeoning state efforts to implement standards-based education in K–12 education (Ewell, 1997). Accreditation organizations, 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 but also far more performance conscious and data hungry than they were three decades ago. Assessment has thus become an unavoidable condition of doing business: Institutions can no more abandon assessment than they can do without a development office.

The last 20 years have also seen a revolution in undergraduate instruction. In part, this results from technology. 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. These 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 are therefore 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, for instance, 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 & 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* foresaw.

Defining the Territory

Although familiar elements of the academic landscape like *teaching*, *courses*, and *degrees* have evolved some reasonably common meanings through continuing use, distinctions among such concepts as *outcomes*, *learning*, *assessment*, and *effectiveness* remain relatively underdeveloped. But discussing learning outcomes sensibly requires an approach that can appropriately distinguish

(a) different levels of analysis, (b) different kinds of "results" of an academic experience, and (c) different perspectives or viewpoints. One way to begin to make sense of this topic conceptually, therefore, is to think systematically about each component of the core concern of student learning outcomes. Doing so first requires discussion of what is meant by an *outcome* and how this is different from other dimensions of activity. Second, it demands distinctions among units of analysis—at minimum, individual students and aggregations of students grouped by characteristic, academic program, or institution. Third, it requires one to distinguish learning from other kinds of "good effects" that students may experience as a result of participating in a postsecondary experience. Fourth, it necessitates explicit consideration of how we know whether (and to what degree) any of these results has occurred, and to what causes we can attribute them. A brief tour of the terminology associated with assessment is provided next, organized around two categories—attributes of collegiate results and attributes of assessment approaches and evidence (key terms are noted in italics).

Attributes of College Results

The definitions that follow are merely the central tendencies of a large and diffuse literature that has evolved over many years. Readers should be aware that some of these terms are defined differently by different authors and some remain contested. But they are nevertheless reasonably consensual across a wide body of practice within the learning outcomes tradition. A glossary is also provided within this volume.

"Outcomes" Versus "Outputs"

Although an *outcome* in current academic usage is clearly the result of institutional and student activities and investments, there is a fair degree of conceptual consensus that not all results are properly considered outcomes. Numbers of graduates, numbers of teaching hours generated, or instances of service or research products are clearly results of what an institution of higher education does. But they are more commonly defined as *outputs* of higher education. An output is the result of program participation that specifies types, levels, and targets of service and is often measured as a count. Other dimensions of institutional or program performance like *efficiency* or *productivity* are equally the results of what an institution does, and assessing them may be important for evaluative purposes. But they are not the same thing as outputs. This latter kind of performance constitutes the central conceptual foundation of what has come to be called *institutional effectiveness*, which examines the extent to which an institution as a whole attains the performance goals that it establishes for itself. Although outputs and performance

are predominantly institution-level concepts, outcomes are only visible at the institutional level by aggregating what happens to individual students. An *outcome*, therefore, can be most broadly defined as something that happens to an individual student (hopefully for the better) as a result of her or his attendance at an institution of higher education and/or participation in a particular course of study. Outcomes are usually measured with a test or an assessment.

Learning as a Special Kind of Outcome

Similarly, relevant and valuable outcomes are not confined to learning because students can benefit from their engagement in postsecondary study in many other ways. Additional *behavioral outcomes* that may result include employment and increased career mobility, enhanced incomes and lifestyles, the opportunity to enroll in more advanced educational studies, or simply a more fulfilled and reflective life. Presumably these are related to learning in some way, and evidence that students have obtained such benefits is often used by institutions as a proxy for instructional effectiveness. But the learning outcomes literature emphasizes that such subsequent experiences should not be confused with actual mastery of what has been taught. Although equally an outcome and frequently examined by institutions, student *satisfaction* with the university experience should also not be confused with learning. Certainly, satisfaction is important—especially if it is related to *motivation* and *persistence* (and therefore continued opportunity to learn). *Student learning outcomes*, then, are properly defined in terms of the particular levels of knowledge, skills, and abilities that a student has attained at the end (or as a result) of her or his engagement in a particular set of teaching/learning experiences.

Learning as "Attainment"

When *learning* is defined in terms of the levels of *attainment* achieved, it requires learning outcomes to be described in very specific terms. Although institutions, disciplines, and professions vary considerably in the ways (and the extent to which) learning outcomes are described, several broad categories are usually distinguished. *Knowledge* or *cognitive outcomes* generally refer to particular areas of disciplinary or professional content that students can recall, explain, relate, and appropriately deploy. *Skills outcomes* generally refer to the learned capacity to do something—for example, think critically, communicate effectively, collaborate productively with colleagues, or perform particular technical procedures—as either an end in itself or a prerequisite for further development. *Attitudinal* or *affective outcomes*, in turn, usually involve changes in beliefs or the development of certain values—for example, empathy, ethical behavior, self-respect, or respect for others. Learned *abilities*

or *proficiencies* typically involve the integration of knowledge, skills, and attitudes in complex ways that require multiple elements of learning. Examples include leadership, teamwork, effective problem-solving, and reflective practice. All such taxonomies require institutions or programs to define *learning goals* or *learning objectives* from the outset as guides for instruction and as benchmarks for judging individual student attainment. Expressed in terms of *competencies*, moreover, such goals describe not only what is to be learned but also the specific levels of performance that students are expected to demonstrate. *Certification* or *mastery*, in turn, implies that these specific levels have actually been attained, generally indicated by a credential or award of some kind. Finally, the relationship among these various outcomes is frequently portrayed through a *qualifications framework*—for example, the Lumina Degree Qualifications Profile (DQP).

"Learning" as Development

In many cases, assessment addresses student learning not just in terms of attainment, but in terms of *growth* or *enhancement*. This construction not only emphasizes the unique contribution of the educational program to current levels of student attainment but also requires some knowledge of what levels of attainment characterized a given student before enrollment. *Value added*, "*before–after*," and *net effects* are terms frequently used to describe such longitudinal ways of looking at development. This perspective, of course, need not be confined to student learning. For example, many educational programs base their claims of effectiveness on things like enhanced income, changes in career, or even increased satisfaction. From the standpoint of quality assurance, both attainment and development may be important. Certification of specific levels of knowledge, skill, or ability for a given program completer—for example, in the form of a licensure examination—is thus intended to guarantee that the certified individual is able to perform competently under a variety of circumstances. Evidence of this kind is claimed as especially important for employers seeking to hire such individuals or the clients who seek their services. Evidence about value added or net effects, in contrast, will be especially important elements of quality for prospective students who are looking for institutions or programs that will benefit them the most, or for policymakers and the public who seek maximum payoff for the resources that they have invested.

But the classic approach to assessing learning gain—testing students on entry, then retesting them at exit—poses perplexing conceptual issues and formidable methodological problems. Conceptually, it can be argued, a pretest is simply silly because students have not yet been exposed to the subject being tested (Warren, 1984). The terminology of *value added*, moreover, suggests a mechanistic view of education in which students are viewed

as products and learning merely additive. Actually determining growth, meanwhile, entails multiplicative sources of measurement error and can sometimes lead to real misinterpretation of underlying phenomena (Baird, 1988; Banta, Lambert, Pike, Schmidhammer, & Schneider, 1987; Hanson, 1988). Because classic test-retest is so difficult to implement in practice, a frequently used alternative is an analytical model based on multivariate statistical controls. Here, a given student's actual performance on an assessment is compared to a "predicted" performance estimated from scores obtained on other assessments (e.g., admissions tests) and additional demographic descriptors.

Attributes of Assessment Approaches and Evidence

Most assessment work involves the collection and use of aggregated information about student abilities (either in absolute or value-added terms). But assessment methods are also used to certify individual students for the award of credentials or credits. Both of these can be examined from the point of view of attainment against established standards (*criterion-referenced assessment*) or from the standpoint of how the performance of an individual or group compares to others' (*norm-referenced assessment*). The term *evaluation* also commonly refers to evidence-gathering processes that are designed to examine program- or institution-level effectiveness. But the object of evaluation usually extends beyond learning outcomes to examine a much wider domain of institutional performance. Finally, all of these applications can be undertaken from a *formative* standpoint (i.e., to advise or improve performance) or from a *summative* standpoint (i.e., to judge performance for a decision or the record).

Validity and Reliability

All assessment measures are in some way flawed and contain an error term that may be known or unknown. In general, the greater the error, the less precise—and therefore less useful—the measure. But the level of precision needed depends on the circumstances in which the measure is applied. Measures are *reliable* when they yield the same results consistently over time or across settings. Assessment measures need to be founded on reliable measurement procedures, but they also need to be designed to operate under less-than-ideal measurement conditions. Uses or interpretations of measures are established as *valid*, in turn, when they yield results that faithfully represent or describe the property or properties that the assessment is intended to measure within a specific context. Note that a measure can be used for multiple purposes. Consequently, validity is a property of the use or interpretation rather than the assessment or test. Historically, four types of validity were

generally considered. *Construct validity* is present when the assessment accurately addresses the ability that it is supposed to address. *Content validity* is present if all the elements of the ability (or domains) are examined and in the proper balance. *Predictive validity* is present if results of the assessment enable confident predictions to be made about future test taker behavior when deploying the ability in question. *Face validity*, although not considered a form of validity in the measurement literature despite the name, is present if the assessment looks like it measures what it is supposed to measure to most observers. More recently, the meaning of *validity* has evolved in high-stakes situations such as accountability and has come to embrace the appropriateness of the evolving uses and interpretations, which is frequently termed *consequential validity* (Messick, 1988).

The latest revisions to the *Standards* reflect the current consensus that validity is a unitary concept that should be established with multiple sources of evidence. The different types of validity evidence are considered under the broad term of *construct validity* to include *content*, *response process*, *structure*, *relations to other measures*, and *consequences* (AERA, APA, & NCME, 2014). This is further discussed in chapters 4 and 6 of this volume.

Types of Assessment Approaches

Assessment approaches and instruments like the ones addressed in this volume are of many kinds and there is no generally accepted taxonomy that embraces them all. Among the most common, however, are the following. *Standardized examinations* typically address cross-cutting attributes such as communications skills, quantitative reasoning, and critical thinking or specific disciplinary content areas. They are "standardized" because their construction allows valid comparisons to be made across performances occurring in different settings (e.g., institution, program). Standardized examinations, in turn, are typically *forced choice* in which a single response set is provided for each item, only one of which is correct. More rarely, they contain *constructed response tasks*, which require the respondent to write essays or short answers of some kind that must be physically scored. Evaluating these requires use of a *rubric* or *scoring guide*, which is typically a matrix of attributes of the work to be scored and a range of scores to be assigned (generally of four to five points for each). Rubrics or scoring guides can also be applied to other pieces of student work (or *artifacts*) such as essays, performances (e.g., a musical performance or medical procedure), oral reports, exhibits, and so on that are generated as special assignments or in regular coursework. Sometimes such artifacts are assembled in a *portfolio* that contains multiple examples drawn at the end of a student's academic career or at different points within it to show growth.

Evidence and Outcomes

Differences in concept and terminology are also apparent when describing the informational results of an assessment. Here, terms like *measurement* and *indicator* are frequently used, implying that legitimate assessment should yield only quantitative results. Measurements, however, are only a special kind of *evidence*, which has come to predominate as the descriptive term for assessment results in quality assurance contexts. Evidence can embrace the results of both quantitative and qualitative approaches to gathering information, both of which can be useful in examining learning. At the same time, the term *evidence* suggests both the context of making and supporting a case and the need to engage in consistent investigations that use multiple sources of information in a mutually reinforcing fashion. But to count as evidence of student learning outcomes, the information collected and presented must go beyond self-reports provided by students and graduates through such means as surveys and interviews or employment placements to include the direct examination of student work or performance.

This brief tour of the territory of student outcomes assessment can only scratch the surface of the myriad terms and techniques that this volume covers. By providing an overview of concepts and terminology, though, it may help readers link these elements around a coherent whole.

Conclusion: A Prime Directive

This volume is intended to provide comprehensive and detailed coverage of tools and approaches to assessing student learning outcomes in higher education. As readers peruse this rich body of material, it is important to continually bear in mind the overriding purpose of assessment: Provide information that will enable faculty, administrators, and student affairs professionals to increase student learning by making changes in policies, curricula, and other institutional programs, and to ensure these changes are actualized through pedagogy and student experience. This is less a method than a mind-set, and it has several relevant dimensions.

First, the animating motives for assessment should reside within institutions and programs themselves. Far too much assessment in higher education is undertaken at the behest of government bodies and accreditors instead of arising from a genuine interest and concern on the part of institutions and their faculties about what is happening to their students (Kuh et al., 2015). Although accountability is important and the tools and techniques addressed in this volume can help institutions and programs discharge this responsibility appropriately, assessment should not be approached with accountability as its sole purpose. This, in turn, implies that the stance of an assessment

should be proactive rather than reactive: The questions that it seeks to answer should be generated by members of an academic community itself, not some outside body.

Second, assessment should always be characterized by a commitment to go wherever the evidence suggests with respect to drawing conclusions or taking action. This may seem obvious, but all too often, expressed conclusions are affected by preconceived notions about how students are experiencing a teaching–learning situation or how faculty members want or expect them to behave. A familiar example is the widespread surprise that greeted NSSE results reporting that most students spend only about half as much time studying and preparing for class than faculty expect (NSSE, 2000).

Third, those practicing assessment in whatever form should bear in mind a number of admonitions about how to proceed, which counteract some of the many myths that the field has encountered over its 30 or more years of evolution. Ironically, these are best expressed as negatives—things that assessment should emphatically not be about. They are as follows:

- *Not "measuring everything that moves."* In part stimulated by accreditors, it is a common misperception that assessment is about gathering as much data as possible about as many things as possible. This can be a recipe for frustration. Rather, the limited assessment energy that an institution or a researcher can devote to the topic should be carefully allocated to questions that are carefully focused, directed toward matters that are considered important, and that will result in genuine improvement.
- *Not just "checking up after the fact."* Assessment is indeed about outcomes, so concentrating attention on end results is appropriate. But if outcomes data are all that are collected, assessors have no information about what caused or occasioned the results obtained, and therefore have no information about how to intervene or improve. Data about educational processes and student experiences should therefore always accompany information about educational outcomes so that causal inferences about what is responsible for the latter can be made.
- *Not searching for "final" answers.* Like the results of any research, conclusions suggested by assessment are always provisional; they may be overturned by later inquiry, and they need to be consistently revisited and tested in the light of newer evidence. More important, the questions raised by each successive cycle of inquiry are at least as important as their results because they sometimes suggest new lines of research and data collection that had not been considered before.

- *Not always being as precise as possible.* Certainly assessment is a science and, as such, its practitioners should strive for accuracy and precision insofar as this is possible. But the real admonition is to be as precise as necessary, given the question being asked and the level of precision required to take action in response. The well-known aphorism about inappropriate method, "measure it with a micrometer, mark it with chalk, and cut it with an axe" captures the difficulty nicely (quoted in Coduto, 2001). Appropriate levels of precision will thus vary by circumstance including how much information is needed to make a particular decision and the urgency of the action to be taken.

- *Not ever expecting to be done.* Finally, assessment should not, under any circumstances, be regarded as a closed enterprise that ends with definitive answers. The process is never entirely finished; rather, assessment is an important part of a continuous improvement cycle (see Figure 1.1). As previously discussed, each answer begets new questions to be pursued, and the resulting spiral of inquiry continues to generate new insights and associated actions. Although this can seem frustrating, it is a fundamental property of scholarship. And scholarship, after all, is what the practice of assessment should model and embody.

Although the topical coverage of this volume is both comprehensive and detailed regarding theory and technique, readers must never forget that the

Figure 1.1. Continuous improvement cycle.
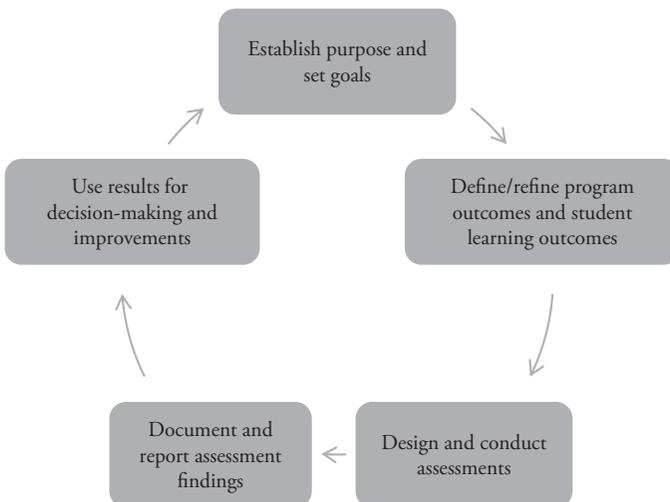
foundational values of assessment lie in action and improvement. As assessment techniques evolve, every assessment approach is a means to an end, and each end is different. Returning to the basic question to be answered or pedagogical problem to be addressed is always a basic prerequisite to effective assessment.

References

Alverno College Faculty. (1979). *Assessment at Alverno College*. Milwaukee, WI: Alverno.

American Association of Higher Education. (1997). *Learning through assessment: A resource guide for higher education*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Association of American Colleges & Universities. (1985). *Integrity in the college curriculum: A report to the academic community*. Washington, DC: Author.

Astin, A. W. (1977). *Four critical years*. San Francisco, CA: Jossey-Bass.

Astin, A. W. (1991). The changing American college student: Implications for educational policy and practice. *Higher Education, 22*(2), 129–143.

Baird, L. L. (1988). Value added: Using student gains as yardsticks of learning. In C. Adelman (Ed.), *Performance and judgment: Essays on principles and practice in the assessment of college student learning* (pp. 205–216). Washington, DC: U.S. Government Printing Office.

Banta, T. W. (1985). Use of outcomes information at the University of Tennessee, Knoxville. In P. T. Ewell (Ed.), *Assessing educational outcomes: New directions for institutional research #47* (pp. 19–32). San Francisco, CA: Jossey-Bass.

Banta, T. W. (1996). The power of a matrix. *Assessment Update*, 8(4), 3–13.

Banta, T. W., & Associates. (1993). *Making a difference: Outcomes of a decade of assessment in higher education*. San Francisco, CA: Jossey-Bass.

Banta, T. W., Lambert, E. W., Pike, G. R., Schmidhammer, J. L., & Schneider, J. A. (1987). Estimated score gain on the ACT COMP exam: Valid tool for institutional assessment? *Research in Higher Education*, 27, 195–217.

Banta, T. W., Lund, J. P., Black, K. E., & Oblander, F. W. (1996). *Assessment in practice*. San Francisco, CA: Jossey-Bass.

Bennett, W. J. (1984). *To reclaim a legacy: A report on the humanities in higher education*. Washington, DC: National Endowment for the Humanities.

Birnbaum, R. (2000). *Management fads in higher education: Where they come from, what they do, why they fail*. San Francisco, CA: Jossey-Bass.

Bowen, H. R. (1977). *Investment in learning*. San Francisco, CA: Jossey-Bass.

Boyer, C. M., Ewell, P. T., Finney, J. E., & Mingle, J. R. (1987, March). Assessment and outcomes measurement: A view from the states. *AAHE Bulletin*, 8–12.

Chickering, A. W. (1969). *Education and identity*. San Francisco, CA: Jossey-Bass.

Coduto, D. P. (2001). *Foundation design: Principles and practices* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

DeZure, D. (Ed.). (2000). *Learning from change: Landmarks in teaching and learning in higher education from Change Magazine 1969–1999*. Washington, DC: American Association of Higher Education.

Eaton, J. S. (2001). Regional accreditation reform: Who is served? *Change: The Magazine of Higher Learning, 33*(2), 38–45.

Enthoven, A. C. (1970). Measures of the outputs of higher education: Some practical suggestions for their development and use. In G. B. Lawrence, G. Weathersby, & V. W. Patterson (Eds.), *Outputs of higher education: Their identification, measurement, and evaluation* (pp. 51–58). Boulder, CO: Western Interstate Commission for Higher Education.

Erwin, T. D. (1991). *Assessing student learning and development*. San Francisco, CA: Jossey-Bass.

Ewell, P. T. (1988). Implementing assessment: Some organizational issues. In T. W. Banta (Ed.), *Implementing outcomes assessment: Promise and perils, new directions for institutional research #59* (pp. 15–28). San Francisco, CA: Jossey-Bass.

Ewell, P. T. (1991). To capture the ineffable: New forms of assessment in higher education. *Review of Research in Education*, *17*, 75–126. Washington, DC: American Educational Research Association.

Ewell, P. T. (1993). The role of states and accreditors in shaping assessment practice. In T. W. Banta & Associates (Eds.), *Making a difference: Outcomes of a decade of assessment in higher education* (pp. 339–356). San Francisco, CA: Jossey-Bass.

Ewell, P. T. (1997). Accountability and assessment in a second decade: New looks or same old story? In AAHE, *Assessing impact: Evidence and action. Presentations from the AAHE Conference on Assessment & Quality* (pp. 7–22). Washington, DC: Author.

Ewell, P. T. (2002). An emerging scholarship: A brief history of assessment. In T. W. Banta & Associates (Eds.), *Building a scholarship of assessment* (pp. 3–25). San Francisco, CA: Jossey-Bass.

Ewell, P. T., Finney, J. E., & Lenth, C. (1990). Filling in the mosaic: The emerging pattern of state-based assessment. *AAHE Bulletin*, *42*, 3–7.

Ewell, P. T., & Jones, D. P. (1986). The costs of assessment. In C. Adelman (Ed.), *Assessment in American higher education* (pp. 33–46). Washington, DC: U.S. Government Printing Office.

Feldman, K. A., & Newcomb, T. M. (1969). *The impact of college on students*. San Francisco, CA: Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. San Francisco, CA: Jossey-Bass.

Hanson, G. R. (1988). Critical issues in the assessment of value added in education. *New Directions for Institutional Research*, No. 59 (Implementing Outcomes Assessment: Promise and Perils), *15*(3), 53–67.

Heffernan, J. M., Hutchings, P., & Marchese, T. J. (1988). *Standardized tests and the purposes of assessment*. Washington, DC: American Association of Higher Education.

Kuh, G. D., Ikenberry, S. O., Jankowski, N. A., Cain, T. R., Ewell, P. T., Hutchings, P., & Kinzie, J. (2015). *Using evidence of student learning to improve higher education*. San Francisco, CA: Jossey-Bass.

Learned, W. S., & Wood, B. D. (1938). *The student and his knowledge*. New York, NY: Carnegie Foundation for the Advancement of Teaching.

Lenning, O. T., Beal, P. E., & Sauer, K. (1980). *Retention and attrition: Evidence for action and research*. Boulder, CO: National Center for Higher Education Management Systems.

Lenning, O. T., Lee, Y. S., Micek, S. S., & Service, A. L. (1977). *A structure for the outcomes and outcomes-related concepts*. Boulder, CO: National Center for Higher Education Management Systems.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University Press.

Loacker, G., Cromwell, L., & O'Brien, K. (1986). Assessment in higher education: To serve the learner. In C. Adelman (ed.), *Assessment in American higher education: Issues and contexts* (pp. 47–62). Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

McClain, C. J. (1984). *In pursuit of degrees with integrity: A value-added approach to undergraduate assessment*. Washington, DC: American Association of State Colleges and Universities.

Mentkowski, M., & Associates. (2000). *Learning that lasts: Integrating learning, development, and performance in college and beyond*. San Francisco, CA: Jossey-Bass.

Messick, S. (1988). *Meaning and values in test validation: The science and ethics of assessment*. Princeton, NJ: Educational Testing Service.

National Governors Association. (1986). *Time for results*. Washington, DC: Author.

National Institute of Education, Study Group on the Conditions of Excellence in American Higher Education. (1984). *Involvement in learning: Realizing the potential of American higher education*. Washington, DC: U.S. Government Printing Office.

National Survey of Student Engagement. (2000). *The NSSE 2000 report: National benchmarks of effective educational practice*. Bloomington: National Survey of Student Engagement, Indiana University.

Nichols, J. O. (1989). *Institutional effectiveness and outcomes assessment implementation on campus: A practitioner's handbook*. New York, NY: Agathon Press.

O'Banion, T. (1997). *A learning college for the 21st century*. Washington, DC: American Council on Education, Oryx Press.

Pace, C. R. (1979). *Measuring the outcomes of college*. San Francisco, CA: Jossey-Bass.

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

Palomba, C. A., & Banta, T. W. (1999). *Assessment essentials*. San Francisco, CA: Jossey-Bass.

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students*. San Francisco, CA: Jossey-Bass.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students, volume 2: A third decade of research*. San Francisco, CA: Jossey-Bass.

Smith, M. K., Bradley, J. L., & Draper, G. F. (1994). *Annotated reference catalog of assessment instruments, catalogs A–G.* Knoxville: Assessment Resource Center, University of Tennessee–Knoxville.

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

Thornton, G. C., & Byham, W. C. (1982). *Assessment centers and managerial performance*. New York, NY: Academic Press.

Tinto, V. (1975, Winter). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*, 89–125.

U.S. Department of Education, National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

Walvoord, B. E., & Anderson, V. J. (1998). *Effective grading: A tool for learning and assessment*. San Francisco, CA: Jossey-Bass.

Warren, J. (1984). The blind alley of value added. *AAHE Bulletin*, *37*(1), 10–13.

Whitaker, U. (1989). *Assessing learning: Standards, principles, and procedures*. Chicago, IL: Council on Adult and Experiential Learning.