A Trout in the Milk

Logistical and Methodological Considerations in the Assessment of Affective Learning Outcomes

Some circumstantial evidence is very strong, as when you find a trout in the milk.
--Henry David Thoreau

Executive Summary

The purpose of this paper is to discuss challenges associated with the assessment of learning outcomes in the affective domain (e.g., attitudes, value orientations, motivations, traits, and so on). The discussion occurs in the context of accreditation criteria as set forth in recent documents from the Higher Learning Commission of the North Central Accreditation Association.

The argument in this paper is that Commission statements that all mission-derived student learning outcomes must be assessed and that faculty should have the fundamental role in doing so pose logistical challenges in the assessment of affective student learning outcomes.

Faculty workloads are heavy, in part because of their responsibility for ongoing assessment of curriculum-based learning outcomes and consequent curriculum and course modifications. In addition many faculty members feel inadequate to assess mission-based affective outcomes, such as “citizenship,” or “respect for diversity.” These outcomes appear “ineffable” and difficult to define, much less assess.

Student affairs professionals, in contrast, deal primarily with the affective dimensions of the college student experience and have built a considerable body of research and theory exploring these dimensions. Therefore, collaboration between student affairs and academic affairs in the assessment of affective outcomes is one way to meet the accreditation challenge.

Most faculty and student affairs professionals have not had adequate training in outcomes assessment, especially in the very difficult assessment of affective learning outcomes. That expertise could be supplied by faculty members or student affairs specialists familiar with research methodologies used in the study of personality traits, attitudes, and values and in comparable methodologies employed in the conduct of outcome studies of the effectiveness of psychotherapy and counseling. These methodologies have been underutilized in higher education outcomes assessment although they have the potential to enrich the meaning and usefulness of outcomes assessment findings.
A Trout in the Milk

Logistical and Methodological Considerations in the Assessment of Affective Learning Outcomes

Some circumstantial evidence is very strong, as when you find a trout in the milk.
--Henry David Thoreau

Part 1: Logistical Considerations

Introduction
In the last decade, the evaluation of quality and effectiveness in higher education has focused on student learning outcomes. An extensive literature has grown around outcomes-based assessment, growth stimulated in part by criteria used by such accrediting associations as the Higher Learning Commission of the North Central Accrediting Association (2003).

Student learning outcomes, it can be argued, provide the clearest indication that an institution is fulfilling its educational mission. If students cannot demonstrate they have learned, it is difficult to say that an educational institution really is, well, educating.

Prior to the emphasis on evidence that students are learning what a college or university says it is teaching, the typical evaluation of educational excellence was based on what might be described as resource indicators. Examples of such indicators include size of endowment, percentage of terminally-degreed faculty, size of library holdings, standardized entrance test scores of freshmen, percent entering students graduating, student-faculty ratio, and so on. Perhaps the most well-known publication using such criteria is America’s Best Colleges, published annually by U.S. News & World Report (cf., for example, America’s Best Colleges, 2004 Edition, 2004)

The limitations of resource indicators as measures of educational quality have been noted for some time (cf., for example, Astin, 1985, 1993). Dissatisfaction with these limitations has led, for example, to a focus on the importance of sound educational practices as an essential precursor to effective learning and an indicator of academic excellence. The National Survey of Student Engagement (NSSE), stimulated by the work of George Kuh and associates at Indiana, is the premier current illustration of this emphasis.

The NSSE assesses student perceptions of five “activity clusters” (e.g., “Student interaction with Faculty Members”) associated with desirable outcomes in college. The authors of the NSSE intend the results obtained from it to serve as ‘national benchmarks of effective educational practice” (2001 NSSE Viewpoint, p. 4). Over the last several years, the NSSE has become increasingly accepted in higher education circles as a useful instrument for the benchmarking purposes intended.

For institutions of higher education accredited by the Higher Learning Commission of the North Central Accrediting Association, however, neither resource indicators of quality nor information about benchmark levels of educational practices are sufficient to demonstrate institutional fulfillment of its stated educational mission. Direct evidence of student learning is required. Indirect measures (such as those mentioned above, and others) can

---

1 The quotation can be found in Schott (2002, p. 114).
supplement and support direct demonstrations of learning, but not substitute for them (Lopez, 1996). Thus it is important to know what is meant by "direct evidence"?

Curriculum-based student learning outcomes are at the center of any college or university’s demonstration that it is fulfilling its educational mission. And because of this centrality, only the best available indicators of learning should qualify as satisfactory evidence. That best indicator is a direct demonstration of learning by students. 2 "To be useful," Lopez argues, “an [educational] objective should be clearly expressed, measurable, and state explicitly what students will be able to do, not describe an internal state” (Lopez, 1998, p. 38).

**Learning outcomes: cognitive, behavioral, affective**

Direct demonstrations of learning by students are necessary and feasible indicators in what are traditionally known as the “cognitive” (knowledge acquisition) and “behavioral” (skill acquisition) domains of learning (Lopez, 1996). Cognitive and behavioral learning outcomes are the stuff of curriculum-based assessment. However, affective learning, which includes changes in values, attitudes, and beliefs, is less often the focus of assessment programs that focus on the academic curriculum.

Unlike cognitive and behavioral learning outcomes, outcomes in the affective realm often resist direct demonstrations of learning that are persuasive, much less conclusive. In this case “indirect measures,” such as surveys of students, alumni, and employers, can “be useful in determining change and growth in what students have gained in the affective domain” (Lopez, 1996, p. 16), although this substitution may not be entirely satisfactory. 3

In the second part of this paper (“Methodological Considerations”, p. 7) the issue of sufficiency in the assessment of affective outcomes is considered at length. But first we look at some of the reasons why affective outcomes are infrequently assessed in faculty-led assessment efforts, reasons that lead to the conclusion that most faculty are ill-prepared to engage in this type of assessment.

**The Faculty meet the ineffable**

Affective learning outcomes are often the stuff of organizational mission statements, where institution-wide student learning outcomes are frequently found. Despite their prominent position, however, these outcomes are often neglected in a higher learning organization’s assessment of student learning. One reason is that they are difficult to define. Hamilton (2003) sums up the matter this way:

---

2 Course grades do not qualify here. Lopez notes that grades, and the GPAs built on individual grades, are NOT indicators of student learning, not even of the indirect kind. Labeling them ‘non-measures,” Lopez (1996, p.15) states that “neither grades nor GPAs are adequate or reliable measures of student learning across an undergraduate major or graduate/professional program of study.”

3 With regard to surveys of students in particular, “student-satisfaction surveys are of little or no value in assessing the incremental learning that a student has acquired during an undergraduate or graduate program. However, pre- and post-surveys that measure changes in attitudes towards values or beliefs yield important information about the attitudinal development of students” (Lopez, 1996, p. 16). In a later paper, Lopez seemed more skeptical of the value of surveys: “One problem with an over-reliance on survey instruments (e.g., student, alumni, employer surveys) is that they yield self-report data, providing only participant’s opinions on what they have learned... “In other words, surveys do not focus on what the student has actually learned, nor do they in any sense measure it” (Lopez, 1998, p.39).
If we were to gather groups of faculty from around the nation, and ask each group to identify what undergraduate education hopes to achieve, most would come up with fairly similar lists: students emerging from higher education with baccalaureate degrees should have knowledge and skills that go beyond the cognitive and that include personal, social, and civic attributes. Expectations for our graduates would include enhancing values and ethics, critical thinking, integrating and applying knowledge, responsible citizenship, appreciating diversity, or aesthetic discernment. Yet we rarely assess improvement and achievement in these areas. We rarely even try to agree on what we mean by them, which is why they are often referred to as “the ineffables of education.” But without these ineffables, of what value is cognitive proficiency (Hamilton, 2003, p. 131).

If faculty endorse, but “rarely even try to agree,” on the meaning of learning outcomes in the affective domain, it is unreasonable to expect them to focus resources on their assessment. Yet, faculty are often assigned the primary responsibility for the assessment of student learning outcomes of all types. Note, for example, the following statement from the Higher Learning Commission’s most recent (2003) Criteria for Accreditation, Commission Statement on Assessment of Student Learning.

The Commission appreciates that effective assessment can take a variety of forms and involve a variety of processes. However, faculty members, with meaningful input from students and strong support from administration and governing board, should have the fundamental role in developing and sustaining systematic assessment of student learning. Their assessment strategy should be informed by the organization’s mission and include explicit public statement regarding the knowledge, skills, and competencies students should possess as a result of completing course and program requirements; it also should document the values, attitudes and behaviors faculty expect students to have developed. (HLC, 2003, p.3.4-2, emphasis added).

The Higher Learning Commission’s twin guiding directives that (1) all mission-related student learning outcomes be assessed and (2) that faculty have “the fundamental role in developing and sustaining systematic assessment of student learning” could well be on a collision course. That is, a commitment to (2)—unless carefully delineated, understood, and implemented—dooms the achievement of (1).

**Faculty expertise and will in the assessment of student learning outcomes**

Without question, the faculty—as teachers, scholars, and guardians of the academic curriculum—are the core human resource for higher learning organizations. Without them the mission of any higher learning organization simply cannot be accomplished, since faculty are essential to the very nature and identity of such organizations. It can’t be simpler: no faculty, no higher learning organization. Period. End of story.

As the core resource in higher learning organizations, faculty—and the academic functions for which they are responsible—are accorded the largest share of whatever financial and physical resources a higher learning organization has available. Faculty, and what they do, are typically held in the highest esteem in such organizations, with all other organizational personnel and enterprises existing primarily to support them and the academic curriculum.

All that said, for most faculty, there is little in their education—undergraduate, graduate, or professional—that provides them with expertise in the assessment of student learning
outcomes, whether curriculum-based or institution-wide. Nor are they expected to have that expertise, either as a requisite for appointment or for promotion.

Academicians are, first and foremost, experts in their academic disciplines, or subfield within the discipline, and become core members of a higher learning organization on that basis. As scholars, faculty are expected to have expertise in their field of study and perhaps advance the intellectual boundaries of their subject. As teachers, faculty are expected to meet with students, engage them in a particular domain of the curriculum, and evaluate their academic achievement in this domain.

If faculty are given primary responsibility for assessment of student learning outcomes, it is reasonable to expect they might do best assessing outcomes that flow directly from expertise in their academic discipline. After all, who better than disciplinary faculty can set the academic goals and objectives of their part of the overall curriculum, and then assess their attainment? Even here, however, most faculty need training in disciplinary-appropriate assessment methodology to carry out satisfactory learning outcomes studies.

Challenges (and time commitments) increase significantly when faculty are asked to take primary responsibility for learning outcomes assessment beyond their disciplinary expertise. The classic case here is assessment of learning outcomes associated with the general education curriculum. General education courses, particularly those offered to juniors or seniors, are often interdisciplinary in nature or intended to be “integrative.” The learning outcomes are, correspondingly, more complex than one might find in many disciplinary courses.

Faculty cannot shirk the responsibility to assess general education learning outcomes, however difficult and time-consuming this enterprise might be. Faculty develop and teach the courses there. Besides, general education is often asserted to be the core or academic backbone of the curriculum as a whole. Its courses affect every student. It is essential to determine how effective this core is in furthering the educational mission of the institution.

In its Commission Statement on General Education, the Higher Learning Commission states its expectation that every higher learning organization “clearly and publicly articulates the purposes, content, and intended learning outcomes of the general education it provides for its students” (HLC, 2003, 3.4-3). The Commission’s additional expectation is that “the organization’s faculty exercises oversight for general education and, working with the administration, regularly assesses its effectiveness against the organization’s stated goals for student learning (HLC, 2003, 3.4-3).

The assessment of complex learning outcomes in general education (or in advanced discipline-based courses) typically requires specialized training for the faculty involved and a commitment of time and energy that, coupled with responsibilities to one’s academic discipline and other faculty duties, can be exhausting. But assessment of general education learning outcomes is clearly a faculty responsibility and they must do it.

The Faculty and affective learning outcomes

In the accreditation guidance provided by the Higher Learning Commission, the assessment work of the faculty is not yet finished. After assessment of the “knowledge, skills, and competencies” in the disciplines and in general education, now to be defined and assessed are the “values, attitudes, and behaviors faculty expect students to have developed” (HLC, 2003, p. 3.4-2).

Can—should—faculty realistically “have the fundamental role in developing and sustaining systematic assessment” (HLC, 2003, p. 3.4-2) of these affective outcomes? Their definition and systematic assessment is steeped in the ineffability noted by Hamilton, above. Further, some faculty question whether they are even amenable to
measurement and assessment at all. So, how can higher learning organizations meet the Higher Learning Commission’s assertion that faculty members should have the primary assessment role in this area?

One possible solution is for faculty to limit their assessment efforts to those affective outcomes they determine are both essential and specifically-aligned with the knowledge and skills that comprise discipline-- or general education-related learning outcomes. Even though most faculty may not be well-equipped to devise and implement assessment of these outcomes, additional training may well be sufficient to help them make a good-faith attempt to do so.

But there are other affective learning outcomes. Those that are mission-based are often organization-wide and quite broad in scope (e.g., “responsible citizenship”). It is unlikely faculty, already burdened, would be willing and able to take primary responsibility for the assessment of those outcomes as they simultaneously address the assessment of curriculum- and discipline-based learning. If they do not assume this responsibility, it is quite possible that mission-based affective outcomes will fall through the assessment cracks and remain at the level of platitudinous “catalog rhetoric”—unless another campus constituency is willing to exert leadership here. The most likely candidate is student affairs.

The role of student affairs in the assessment of affective outcomes

By virtue of their education and training, student affairs professionals are prepared to roam the realm of affective outcomes in relative comfort. In their implementation of the cocurriculum, they are supported by decades of experience and—more importantly—by a body of theory (e.g., Chickering and Reisser, 1993; Evans et al., 1998, Giddan and Price, 1985; Whiteley et al., 1982), and a corpus of research findings focused on the (affective) lives of college students and the personal development of young adults (see Astin, 1993 and Pascarella and Terenzini, 1991 for compendia).

Given the above, student affairs professionals represent a valuable institutional resource in the assessment of mission-based affective learning outcomes. It makes sense that they should partner with interested faculty in this assessment. And it is reasonable to expect them to take, as a minimum, some portion of the responsibility for this assessment, whether or not faculty are available for collaboration.

Although student affairs professionals have an understanding of student affective development and use the cocurriculum to further it, that does not necessarily mean they have expertise in developmental outcomes assessment. As is the case with faculty, the majority of student affairs personnel have not had extensive training in outcomes assessment methodologies or their effective implementation. Therefore, they are just as much in need of support and training as their colleagues in the professorate. Without this support and training, the alliance of academic and student affairs in the assessment of affective outcomes may illustrate the cliché of “the blind leading the blind.”

Auden wryly echoes this skepticism with the poetic admonition that…Thou shalt not sit/With statisticians nor commit/A social science.

See Walvoord (2004), pp. 99-103 for an assessment paradigm entitled “University-wide Measures to Assess University-wide Learning Goals” that specifically includes student affairs in the assessment of “social and moral” responsibilities.

The majority, but not all. See, for example, the work of Astin (1991), Bresciani et al., 2004, Schuh and Upcraft (2001), and Upcraft and Schuh (1996).
Part 2: Methodological Considerations

Challenges involved in the assessment of curriculum-based cognitive and behavioral learning outcomes are substantial, but appear more manageable when compared with similar challenges associated with the assessment of mission-based affective outcomes. Here’s why.

Curriculum-based learning outcomes, especially those within an academic discipline, are focused on knowledge and skills that are often relatively specific. The specificity makes precise definitions of outcomes more likely. The precise definition, in turn, makes it easier to arrive at an assessment methodology that is understandable and readily linked to the outcome. The whole process is very likely to “makes sense” to anyone willing to take a look at it.

As learning outcomes move from the specific to the more general (as they are likely to do in advanced disciplinary or interdisciplinary courses) problems arise. Consensually-validated definitions of general outcomes are harder to come by, and methods of assessment that meaningfully demonstrate that learning has taken place are more difficult to devise and implement.

However, there is lots of help available to address these problems. The assessment of curriculum-based learning outcomes has been the subject of intense interest for well over a decade. As a result, there is a significant body of knowledge about “what works.” Illustrations of best practices are presented at numerous assessment conferences and are widely available in print for institutions that cannot afford to send their assessors to such gatherings. Consultants travel throughout academia, offering on-site practical training in outcomes assessment to institutions preparing for accreditation visits. As a result, there is usually no need to reinvent the outcomes assessment wheel (although some institution-specific tweaking may be necessary).

The culture of outcomes assessment in higher education is focused on demonstrations (“direct evidence”) of cognitive (knowledge) and behavioral (skills) outcomes at the curricular level. This culture is influenced by the guidance offered by accrediting associations, guidance which emphasizes outcomes assessment of curriculum-based learning by faculty. When reading published accreditation criteria, it is difficult to find many statements emphasizing the necessity of assessing affective learning (or to the importance of such learning itself) even though affective outcomes are most often embedded in the organization’s mission statements and touted by institutional representatives as core characteristics of its graduates. On that basis alone, one might argue they deserve more prominence in both the higher education assessment literature and in published criteria for accreditation.7

Aside from their relative neglect by higher education assessment, there is one more reason why assessment of affective learning lags behind—the very nature of affective outcomes themselves. When compared with cognitive and behavioral learning, outcomes in the affective realm have features that, well, do make them appear ineffable.

---

7 Two assessment position papers—one from an academic affairs perspective, the other from student affairs—which are holistic and comprehensive in scope: Nine principles of good practice for assessing student learning (AAHE, 2003) and The student learning imperative (ACPA, 1996). The recent Learning Reconsidered (Keeling, 2004) is another from the student affairs perspective. Powerful Partnerships: A shared responsibility for learning (1998) published jointly by AAHE, ACPA, and NASPA is the most detailed, with illustrative program examples.
**Fishing for the Ineffable**

The assessment of affective learning in higher education has, as noted earlier, relied heavily on student self-reports from standardized or institution-developed surveys. The only direct measure that appears in wide use is the student essay (or comparable document) evaluated by trained readers using mutually-agreed-upon criteria (Walvoord, 2004, pp. 3-4). The combination of survey data, criteria-rated essay, and (perhaps) self-reported tally of relevant behaviors is state of the art in the assessment of “goals that faculty and institutions hold most dear...[goals that] may be the most important qualities that higher education can nurture in the citizens of the future” (Walvoord, 2004, p. 3).

Given the centrality and importance of mission-based affective learning outcomes, it is appropriate to expand this assessment paradigm by using the expertise psychology in particular has built over the last seven decades in the measurement of the affective dimensions of human nature.

These dimensions—the emotions, desires, needs, attitudes, values, goals, traits and other “personal” qualities that contribute to the making of a human being—are subjects of scientific study in the social sciences, most particularly by several branches of psychology. These branches include social psychology, personality, and psychometrics. The human services professions, such as psychiatric social work, counseling and guidance, psychotherapy, teacher education, and personnel management are often the end users of theoretical advances and research findings from the branches of psychology just-listed.

In higher education, student affairs professionals are the chief consumers of this psychological literature, at least as it relates to an understanding of the personal development of young adults. In general, student affairs is focused on the application of knowledge to direct service, just as the human services professions mentioned above. As a result, their knowledge of research methodologies in the behavioral sciences often is not extensive enough to equip them to conduct research, although they are sympathetic to such efforts and apply what findings they can to the cocurriculum.\(^8\)

This sympathy makes for a relatively easy alliance between the branches of academic psychology mentioned above and student affairs. True, the alliance must bridge the political factors that contribute to the separation of faculty from administrative staff in most higher education organizations, but it is feasible, a feasibility made possible by mutual interest in the affective development of young adults and its relation to the college experience.

Psychology is not the only academic discipline that can make common cause with student affairs in the assessment of affective learning. Sociology and political science are two others, as well might be philosophy and religious studies. But this paper will focus on the contributions of psychology because it is psychology that has developed the methodological sophistication needed to tackle the difficult task of defining and measuring the affective dimensions of the human person. And it is psychology that has applied similar methodology to assessment of the seemingly intangible outcomes of psychotherapy and personal counseling, human services that intend to “educate” for effective living.

\(^8\) Like their faculty colleagues, however, student affairs professionals can employ assessment techniques which require minimal technical sophistication, such as criteria-based rubrics and student portfolios. These techniques are widely used in the assessment of knowledge and skills at the program level. Bresciani et al (2004) provide student affairs staff with an excellent primer on how to use these—and related—approaches.
A trout in the nomological net

A nomological network, as described in a paper by Cronbach and Meehl (1955), refers to a model for relating constructs or abstract concepts (“socially aware and responsible individual,” say, or “independently creative person”) to empirical realities. The model is “nomological” because it seeks to state the “lawful” relationships (1) among abstract concepts, (2) among observable properties of these concepts, and (3) between the concepts and the properties which illuminate them.

The general idea of a nomological network is what we want to use here. That idea is to weave a net of evidence around ill-defined affective outcomes and thereby gradually capture their meaning. The “strands” of the net are the relationships between and among the affective outcomes of interest and their observable properties as exemplified through a variety of assessment methods (e.g., criteria-rated essays, self-report surveys, ratings by various associates, skill or knowledge demonstrations, individual or group interviews, self-reported enumeration of behaviors assumed to be related to the outcome, etc.).

This approach to affective outcomes assessment is an informal adaptation of a research methodology to assess the validity of abstract psychological concepts called the multitrait-multimethod matrix (Campbell & Fiske, 1959). Applied to learning outcomes assessment, the basic idea is to assess the statistical relationships (correlations) among the results obtained by different approaches to the assessment of a learning outcome in order to better understand the meaning of the outcome itself. As the network of relationships is extended and elaborated, the circumstantial evidence surrounding the complex learning outcome of interest becomes “very strong” indeed--strong enough to yield valid and useful information on student learning and provide a reasonable basis for institutional improvement comparable to “direct” evidence.

Since publication of the influential papers on the nomological network and multitrait-multimethod matrix, psychologists have developed an array of methods and statistical procedures designed to elucidate and validate the meaning of seemingly vague (yes, ineffable) concepts not amenable to direct demonstration or unambiguous observation. The methodologies and procedures (collectively gathered in the field of study called “psychometrics” or, more informally, “measurement”) have increased in sophistication and technological complexity in the decades since the Cronbach & Meehl/ Campbell and Fiske papers (see Nunnally and Bernstein, 1994, for a comprehensive overview). Faculty and student affairs professionals with appropriate training in psychometrics can make a unique contribution to the assessment of complex student learning outcomes.

Putting the ineffable in context
The contributions of psychology to the assessment of affective outcomes is not limited to methodological assistance. Beginning with the work of psychologist Gordon Allport in the 1920s (Allport, 1921, 1937, 1961), a number of psychologists have focused their research attention on the “normal person.” Maslow’s study of the “self-actualizing person” (Maslow, 1968, 1970, 1971) is probably the most widely-known effort in this area, but there have been many others, so the literature in the field of personality is extensive indeed.

We limit this discussion to recent developments in the field of personality study of particular relevance to affective outcomes assessment in higher education. The broad mission-based affective outcomes at the college and university level can often be subsumed under the concept of “character.” Character can be defined as those aspects of

---

9 It should be noted there are critics within psychology who have questioned the validity and usefulness of current psychometric methodology in the assessment of personality. See Kline (1998) for an example of a sophisticated critique.
a person which are associated with “moral constitution,” or “moral strength.” Colloquially, character is ‘personality with a moral twist.’ Mission-based affective outcomes which exemplify character include “responsible citizenship,” “acting with integrity,” “dedication to human welfare,” “converting to society,” “developing relationships based on mutual respect,” “embracing diversity,” and so on.

The subfield of personality theory and research that focuses on character is called positive psychology, a decade-old attempt to “reclaim the study of character and virtue as legitimate topics of psychological inquiry and informed societal discourse” by focusing on “what is right about people and specifically about the strengths of character that make the good life possible” (Peterson & Seligman, 2004, pp. 3,4). Although seemingly a recent development, positive psychology can trace its origins to the early work of psychologists such as Allport and Maslow, and subsequent studies by a number of behavioral science researchers.

The possible contributions of positive psychology to the assessment of mission-based affective outcomes are several. First, the recent classification/discussion of character strengths by Peterson and Seligman (2004) provides a convenient, resource-rich context in which to embed similar mission-based affective outcomes.

Second, research related to positive psychology can help jumpstart a higher learning organization’s assessment of mission-based qualities of character by providing a base of empirical findings relevant to the qualities of interest. This base can provide the beginning of nomological networks for those qualities.

Third, positive psychology’s focus on character strengths and their relationship to a meaningful life can appeal to members of both C. P. Snow’s two cultures, albeit for different reasons. Humanities faculty may find the derivation of the “virtues” discussed by Peterson and Seligman (2004) appealing, since they come from a cross-cultural review of both Eastern and Western ethical traditions—an acknowledgement of the central roles of philosophy and religion in the civilizing of society. Those of a more scientific bent may be attracted to the empirical manner in which virtue and character are further explored. The broad appeal of positive psychology’s focus on character can thus become a force to help unify commitment to mission-based affective outcomes.

Methodological tools and a theoretical framework are not the only contributions behavioral science can make to the assessment of mission-based affective outcomes. Experience with actual assessment of affective outcomes is another. That experience comes from research on the outcomes of psychotherapy.

**Does higher education need psychotherapy?**

In 1961 the British psychologist, H. J. Eysenck wrote:

> To judge by their writings, some advocates of psychotherapy appear to take an attitude similar to that adopted by Galen, the father of modern medicine, in his advocacy of the wondrous powers of Samian clay: “All who drink this remedy recover in a short time, except those whom it does not help, who all die and have no relief from any other medicine” (Eysenck, 1961, p. 697).

---


Eysenck’s caustic remarks suggest vexation with the relative lack of empirical support for a profession that had assumed the efficacy of its therapeutic interventions and felt little need to support its assumed benefits with anything but occasional anecdotes and expressions of beneficent intentions. Sound familiar?

Attacks such as Eysenck’s on the scientific credibility of psychotherapeutic claims resulted in a gradual focus on therapeutic outcomes that were measurable and verifiable (e.g., reduction in psychiatric symptoms targeted for intervention). In the 1980s and 1990s, psychotherapists sought professional respectability (and financial reimbursement) from health insurance companies as “personal pay” clients or patient became fewer in number. Respectability and financial remuneration came at a price, however. The cost was compliance with recommended standards of practice imposed by licensing and professional organizations and reinforced by insurance companies. Sound even more familiar?

The standards of practice and related treatment guidelines did not arise in an informational vacuum. Rather, they were based, at least in part, on findings from research studies on the processes and outcomes of actual psychotherapy “in the field” and on results from treatment analogs studied in controlled settings such as university-based psychological laboratories and clinics. The findings suggested treatment guidelines; the treatment guidelines, in turn, led to new research studies with the intention of demonstrating either better efficacy or improved efficiency for a particular type of therapy (read: educational practice) in the treatment of a particular kind of disorder or symptom (read: deficiency or weakness in need of improvement or development). In other words, research on the outcomes of psychotherapy exemplifies the strategy of assessing outcomes and then using assessment findings to improve practice—“closing the loop.”

Similarity is not identity, but…

Similarities between psychotherapy and higher education educational practices should not be pushed too far since there are obvious major differences between the two. But they are worth noting, since what has occurred with the profession of psychotherapy (and the allied fields of personal counseling and psychiatric social work, as well as elementary education) can identify similar challenges and suggested responses in the assessment of affective learning outcomes in higher education. First the challenges.

Despite some forty years of research on the efficacy of psychotherapy, with definite progress that leaves behind the psychotherapeutic version of Samian clay (see Nathan & Gorman, 1998 for a progress report), there is still much that is not known about the relationships between particular interventions and specific outcomes. Conflicting research results are in need of resolution, and there remains a tension between findings from controlled “pure” therapeutic protocols and results from therapy conducted “in the trenches” of everyday practice.

What do these challenges imply for the assessment of mission-based affective outcomes in higher education? The colloquial response is, “Don’t expect miracles.” The professional response is, “Expect slow progress—albeit still progress—when assessing affective outcomes.” The job is complex and multifaceted. As noted above, mission-based affective outcomes resist simple measurement. Making meaningful connections between outcome and educational intervention(s) is even more difficult than measuring an outcome itself. But both are achievable.

13 See Ogles et al. (2002) for a primer on outcomes assessment in psychotherapy.

Circumstantial evidence can be very strong—and sufficient

There is no escape. In the assessment of affective outcomes, self-reports of internal states, of personal demographics, and of past or current behaviors not readily verified by external means provide most of the raw data for analysis. Direct observable demonstrations of knowledge or skills presumed to be associated with a particular affective outcome are sometimes possible (and always very desirable). But such direct demonstrations can be difficult to devise and prohibitively expensive in time or other institutional resources. And, even when possible, direct demonstrations do not escape the same empirical requirements regarding reliability and validity that self-reported information must meet.15

These are lessons from almost a century of empirical study of the human person. Despite the theoretical attraction of empirically grounding such a study in observable behavioral demonstrations of hypothesized traits, values, motives, mood states, and attitudes, it has rarely been possible to do so. Even such a committed and energetic personality research psychologist such as R. B. Cattell (1973, 1979), whose work spanned decades and resulted in literally hundreds of publications, made little progress in his attempts to devise valid and reliable “objective tests” of personality variables—this despite intensive efforts.

So, in the study of affective outcomes self-report, indirect—circumstantial—evidence, is usually the best we have and it must suffice. It can be made “very strong” by continuing to conduct basic research on variables that affect the validity of self-reports themselves (cf. Stone et al., 2000) and, in the field, by employing assessment strategies informed by the ideas behind the nomological network and multitrait-multimethod matrix.

Here’s a quick strategy:

To the extent feasible, define an affective outcome through a combination of self-ratings, self-reported behaviors, ratings by informed others, tallies of activities presumably related to the outcome, focus group or interview data, and direct demonstrations when possible. Follow with an exploration of the relationships among them as a way to clarify the meaning of the affective outcome in question. Then explore the relationships between the clarified outcome and institutional demographics and practices to see if there is at least suggestive support for mission impact. Ideally, take into account student characteristics at matriculation on the outcome (Astin, 1991) to statistically control for their influence. Use results to inform, strengthen (and perhaps reform) relevant institutional practices. Then start assessment cycle over again. Fun, rewarding, exhausting—necessary.

And in conclusion…

Assessing affective outcomes—particularly those that are mission-based and institution-wide—and linking them to institutional practices is more difficult than doing the same assessment of their cognitive and behavioral counterparts. This is not work for folks who prefer quick results or who like their intellectual desks neat and clean. It is indeed a dirty job, but somebody’s got to do it. That “somebody” is most likely a dedicated team of student affairs professionals and faculty members who will need the long term support of a college or university that is proud of its mission-based learning outcomes, takes its identity from them, and is willing to be judged by how well its graduates exemplify them.

One might be proud to hold membership in such an academic community.

15 See Appendix A for a review of reliability and validity and their relationship to assessment.
Bibliography for A Trout in the Milk


Appendix

*Reliability and Validity: What they are and their relevance to student learning outcomes assessment*

All assessment involves measurement. For any measurement to be useful, it must have two characteristics: reliability and validity.

Assessment of psychological characteristics is imperfect. Therefore, measurements of these characteristics will also be imperfectly reliable and valid. However, to be useful in decision-making, measurements must demonstrate some *minimal* level of reliability and validity. Measurements below these minimal thresholds are untrustworthy and may be worse than “educated guesses” or “common sense” because they can be even more misleading. Therefore, it is important to determine the reliability and validity of assessment measures to the extent possible.

**Reliability**

In general, “reliability” refers to “stability.” The most obvious type of stability is temporal: does a particular measure administered under (virtually) identical conditions at different points of time yield (virtually) the same results?

Temporal stability is relative, however. We would not expect moods to be particularly stable over time (by definition). And we would be disappointed if results of measures of academic skills (e.g., “reasoning to a conclusion”) remained the same from freshman to senior year.

So temporal stability is not the same as “permanent.” The stability has to be “reasonable.” A “reasonable” time period is one in which we assume no real changes in the characteristic measured are likely to occur. We would not expect the academic skill of “reasoning to a conclusion” to change much (except for random fluctuations) in the span of a few weeks during which no training on this skill occurred, for example.

Another type of reliability relevant to assessment is called “internal consistency.” Internal consistency may be important if the measuring device is comprised of more than one component (or item). It would be desirable, for example, for a ten item measure of “quantitative reasoning” to have all ten items be primarily indicators of one thing rather than several things (IQ, for example, or reading comprehension). In this case, we would like the measure to be stable “item to item,” with strong relationships among the items and with the “total score.”

Because most psychological characteristics are connected to one or more other characteristics, perfect internal consistency is impossible (and perhaps is not even desirable)—many measures intended to measure a single characteristic may also unintentionally measure others. The goal, of course, is to have a measure that *primarily* assesses what is intended and that is also relatively impervious to the impact of random extraneous factors not relevant to one’s purpose (e.g., lighting in the testing room, time of day, amount of sleep obtained by respondents, etc.).

Unintended variables and extraneous elements are, collectively, labeled “error.” An individual’s (observed) result on a measure is comprised of a “true” score (the one we want) plus “error.” The goal is to have a measure with as much of the former as possible and as little of the latter.
Although the discussion thus far has focused on measures that relate to “tests” and related measures, such as questionnaires, there is another type of reliability very important to ratings made by judges (as in the case of rubrics or portfolio analysis, for example). This kind of reliability is called \textit{inter-rater reliability}. If there is little agreement among Judges A, B, C, and D on their ratings of a particular piece of student work, then inter-rater reliability will be low.

Reliability is fundamental to making sense of assessment findings. If a measurement device or method does not provide reasonably reliable (stable) results, it is not possible to be reasonably \textit{valid} either. A rubber ruler that provides randomly different results each time it is used is not a valid tool for cutting wood. That is, it doesn’t measure what we intend it to measure—it is measuring something else (in this case, perhaps, “rubberiness” rather than a dimension of the wood).

Thus reliability (stability of measurement) is the necessary precursor to validity. Reliability does not, however, appear to be a focus of discussion in higher education assessment and frequently is assumed rather than actually evaluated.

\textbf{Validity}

Simply put, this is the definition of validity: that an assessment tool or methodology measures what we intend—and say—it measures. To the extent it does so, the results are valid. Sounds simple, but it’s not. Validity is multifaceted and somewhat confusing.

There is more than one type of validity. In fact, there are two main types, each with subtypes.

1. First, let’s look at the more easily-understood of the two main types, labeled by Trochim \textit{translation validity}. The term suggests we translate our measurement purpose (to assess, for example, the outcome of “responsible citizenship”) into some operation (an essay with rubric, say, or a test, on dimensions of “responsible citizenship”).

2. We can assert translation validity by saying the measure appears valid on its face (literally, “face validity”). You know, “If it looks like a duck,...” More rigorously, we can assert the measure has “content validity” because several experts (those who are qualified by education or experience to know such things) say the measure covers the essential dimensions of the concept we are attempting to assess.

3. Establishing face and/or content validity is typically the point at which most assessment programs in higher education terminate validation. That may be practical because there’s lots more work ahead if one tries to move on to the second type of validity, \textit{criteria-oriented validity} (see Trochim at website above).

4. Criteria-oriented validity is pretty much what it says: One compares the results of an assessment operation to some \textit{external} criterion (or criteria). That’s confusing, because a favorite outcomes assessment operation—the use of rubrics—requires that criteria be established too. The difference is that the criteria in rubrics are essentially the definition of the concept or construct we are trying to assess (e.g., the components of “responsible citizenship”) whereas the criteria discussed here refer to something external to the definition. Hmmm, like what?

5. Well, if one has already established criteria for a rubric, these criteria become the defining characteristics of the outcome. In effect, we have defined the construct of, say, “responsible citizenship.” Now that we have some notion of what “responsible

\footnote{See William Trochim’s website (accessed at: \url{http://www.socialresearchmethods.net}) for helpful discussions of these—and related—topics.}
”is all about, we can hypothesize what the relationships might be—theoretically—between the various dimensions of the construct and other appropriate phenomena (external criteria).

We might, for example, predict that students with higher levels of assessed responsible citizenship volunteer more often for service projects (predictive validity) or that these same students are more likely to major in one of the social sciences than in one of the physical sciences (concurrent validity).

We could go even further by demonstrating that “responsible citizenship” is more similar to (correlates more closely with) the concept of “embraces diversity” (convergent validity) than to the concept of “understands relationships in the natural world” (discriminant validity). If we do all these things, we have addressed the four subtypes of criteria-related validity. And we have become more knowledgeable about just what “responsible citizenship” is.

**The importance of criteria-related validity**

Why bother with criteria-related validity? There’s at least one really good reason and another just good reason. The really good reason is that assessment programs which do not address criteria-related validity have not fulfilled their primary purpose. That purpose is to demonstrate that relevant components of the institution are making significant contributions to its mission-related student learning outcomes. Information about face or content validity alone won’t fulfill that purpose. Strictly speaking, all they show is “where students are at” without clearly indicating how or why the students “got there.”

We need to make predictions (or, in some cases, “postdictions”) along the lines of, “Students who complete the course, *Christian Service—Its Nature and Value*, will (other relevant variables being equal) score higher on “responsible citizenship” than similar students who do not take the course.” We would want to make that prediction if becoming a more responsible citizen was one of the intended outcomes of the course—one way the course contributed to the institutional mission-based objective of “citizenship.”

Demonstrating the institution is actually doing what it intends is the really good reason to go beyond face and content validity to criteria-related validation. The just good reason is not quite as important, but it is still important: criteria-related validation enriches our understanding of broad mission-based outcomes. This increased understanding may suggest new ways for the institution to impact students or, perhaps, lead to the discovery of positive impacts previously-unknown but already present.

If the increased understanding is shared through oral or written publication with interested higher learning organizations addressing similar outcomes, then one is being collegial in all the best senses. Such publication also makes a contribution to the scholarship of assessment and helps to make “best practices” even better.

There are two “challenges” to conducting criteria-related validation studies. First, they are time- and resource-consuming and they require some methodological expertise. Second, validation is never-ending, one useful finding or discovery typically leading to more than one new question or puzzle. Those conducting validation studies need to like learning.