Change
THE MAGAZINE OF HIGHER LEARNING

JANUARY-FEBRUARY 2007

Counting and Recounting: Assessment and the Quest for Accountability

by Lee S. Shulman

When my daughter Dina returned from her first class in managerial accounting early in her MBA program, I innocently asked how it had gone. I fully expected her to describe her boredom with the rigors of accounting, since pursuing an MBA was decidedly an afterthought for my iconoclastic daughter, who already held degrees in theatre and social work.

Imagine my surprise when Dina responded that accounting was unexpectedly interesting because, she now realized, it should be understood as a form of narrative, a kind of drama. Within the ethical and technical rules of the field, the task of the accountant is to figure out which of the stories of the company should be told through the medium of its "books." Accounting is basically about creating the plot, characters, and setting of the story. As the instructor explained to the class, "Your task is to render an account: to tell the facts of the case, the story of the condition of a company in an accurate and yet ultimately persuasive way."

I was reminded of this conversation as I read through the successive drafts of the Spellings Commission report, with its persistent refrain that higher education must become more accountable, more transparent, and more open to the scrutiny of its stakeholders. The key word is always "accountability," to which the canonical reaction among educators is a reaffirmation of the remarkable diversity of American colleges and universities and the dangers that accompany the specter of standardized testing and a "one-size-fits-all" approach to assessing the quality of a college education.

In the world of business, an account is a story told in quantitative form. It publicly documents all the income and investments that enter the company and all the products and liabilities that emerge from it, all its assets and debits, all its profits and losses. When the books balance, the account is closed: The story has been told.

Indeed, historian of science Mary Poovey argues in A History of the Modern Fact that a significant source for the modern conception of a scientific fact—that which is measurable, replicable, visible, quantitative, and credible—is the invention of double-entry bookkeeping in late-16th century England. Thus accounting was a source for modern scientific conceptions of evidence; then, in full-circle fashion, scientific doctrines became the basis for our contemporary conceptions of account-ability in education.

When I draw our attention, as Dina did mine, to the ghosts of narrative and story-telling that stand behind the counting, measuring, and computations that lie at the heart of modern assessment in the service of accountability, I do not aim to undermine the credibility of assessment. I am not referring to "mere storytelling" as if narrative is a lesser form of discourse. The connections between counting and recounting are built into the etymology of these words in many languages. Thus, in German, to count is zaehlen and to tell (a story) is erzaehlen. Even in Hebrew, a language with utterly different roots than English or German, the verb for counting is l'spor, while the word for telling is l'saper.

I believe the lesson is clear. How and what we choose to count and the manner in which we array and display our accounts is a form of narrative—legitimately, necessarily, and inevitably.

Tools for Counting and Recounting

When Benjamin Bloom led a group of university examiners in the development of the taxonomy of educational objectives in the late 1940s and early 1950s, their goal was to provide a structure within which assessors could determine which story they wished to tell about the learning of their institution's students. They had determined that most of the instruments then in use to assess

students—and thus to render them, their teachers, and their colleges accountable—were exclusively stories of the acquisition and retention of knowledge, of the students' success in recalling facts, events, principles, and concepts they had learned in class or read in their textbooks. Bloom and his colleagues argued that this was an impoverished story, one that missed the most important aspects of the account the examiners needed to give of students' learning.

By elaborating the cognitive outcomes of education into a taxonomy comprised of six categories—ranging from knowledge and comprehension through application, analysis, synthesis, and evaluation—Bloom and his colleagues developed a much richer array of plots and themes for the story of academic performance. A program that appeared to be achieving great success when knowledge alone was measured might look much less impressive if the "higher-order" processes were accounted for. Bloom and his associates also were committed to extending the story from the cognitive to the affective domains in order to include the development of emotions, motivations, passions, and identity.

The power of Bloom's approach to make visible important aspects of learning that would otherwise remain hidden (or to point out their absence) is nicely illustrated by a painful episode in my own history as a learner. When I was an undergraduate at the University of Chicago in the late 1950s, I attempted to cram for the end-of-year comprehensive examination in the history of western civilization—a nine-hour multiple-choice and essay test. I thought I had done quite well on the exam and was thus shocked to receive a "C" for the course. I asked to meet with a member of the Evaluation Office to learn why I had performed so poorly. We sat down and examined my performance, using Bloom's taxonomy as a template. I had "aced" the multiple-choice section, with its emphasis on recall; cramming can be a pretty good strategy for remembering facts and ideas, at least over the short term. But I had simply not studied well enough to integrate the ideas and to be able to synthesize new interpretations and arguments using the knowledge I had crammed into my head.

Had the accounting been limited to a factual knowledge of history, the Shulman narrative would have been one about a highly accomplished student of history. But the richer plot afforded by the design of this assessment told a more complex and less comforting story: Shulman knew the facts of history well but had not yet learned to use them in the service of new ideas or to solve novel problems.

Narratives are enriched not only by changes in plot and theme; introducing new characters as protagonists also has a profound effect. Thus, if the narrative were to examine the learning of discrete sub-groups of students, its complexity and nuance would increase. Is this an institution where students of one particular ethnic background score well across the categories while others do well only in knowledge acquisition but not in the higher-order achievements? Or is this a college where those majoring in the sciences flourish while those studying the humanities flounder? Each of these is a legitimate, "true," and reasonable account—on which the school's accountability will rest. Numbers may offer an illusion of irreducible accuracy and credibility, but they can only be interpreted in the context of the narrative selected and, indeed, the narrative not taken.

The story told by an assessment is thus ultimately a function of the dimensions of measurement that determine the possible directions the narrative might take. So accountability requires that we take responsibility for the story we commit ourselves to telling. We must make public the rationale for choosing that story as opposed to alternative narratives. This requires that we first deliberate with our colleagues and stakeholders about the goals we set, the missions of our schools, and the elaborated conceptions of our purposes.

Only then should we defend the adequacy of the forms of measurement and documentation we employ to warrant the narratives we offer. In the case of educational accountability, we are limited in our recountings by the instruments we use to count. As my colleague Lloyd Bond regularly reminds me, "Since we can't normally measure everything that counts, we can be sure that what will count is what we choose to measure." Taxonomies and indicators are critical aspects of how and with what coherence and credibility these stories can be told.

We can readily see the narrative possibilities for these accounts by examining some of the instruments and indicators that the Spellings Commission singled out. The Collegiate Learning Assessment (CLA) has received a great deal of attention recently and is described in some detail by Richard Shavelson in this issue of *Change*. What story does the CLA tell? The broad domains of its account are critical thinking, analytical reasoning, problem-solving, and writing. The heart of the narrative is the value added by a college education to the educational outcomes of students, rather than the absolute levels they achieve. It chronicles the

development of their learning, thinking, judgment, and communication skills and does not aim or claim to assess domain-specific knowledge, skills, values, or appreciations. Thus, students' performance on the CLA does not correlate with their majors. It is currently used to tell a story about institutions, not individual students.

The National Survey of Student Engagement (NSSE) tells a very different kind of story. Although the items are designed to serve as proxies for outcomes, the instrument itself measures the kinds of experiences students have over the course of their academic careers. While the CLA looks for changes in the performance of students, the NSSE is more attuned to the opportunities the institutions offer and the advantage the students take of them. The NSSE describes institutions in terms of their level of academic challenge; the opportunities they provide for active and collaborative learning; the extent and quality of students' interactions with faculty; the availability and access to enriching extra-curricular experiences; and the extent to which the campus offers a supportive environment for learning and student development.

It's no accident that so many institutions (more than 970 for NSSE and 250 for the CLA) have opted to use one or both of these instruments. Each offers a very different narrative of educational opportunities and accomplishments. While they were not designed to fit together elegantly, they do offer different perspectives on this question: What account can be given of this institution's contribution to the education of its students? Notice, however, that neither instrument tells us anything about the discipline-specific aspects of learning. Do students learn to think like historians? Do they learn to reason quantitatively? Do they come to know the fundamental concepts of science and technology that are needed in the 21st century economy?

The Educational Testing Service's Measure of Academic Proficiency and Progress (MAPP), another instrument specifically identified by the Spellings Commission, attempts to tell a story that gets at some of these differences. Its chapter headings are "Reading," "Writing," "Critical Thinking," "Mathematics," "Humanities," "Social Sciences," "Natural Sciences," and—naturally—a total score. But before we leap to the conclusion that in the MAPP we now have a comprehensive, domain-specific map of student learning over time, we must note that the long form of the assessment takes all of two hours and includes 108 items, which is rather sparse for a substantive evaluation across so many areas. And ETS also offers an abbreviated form of the MAPP that contains only 36 items and can be administered in a total of 40 minutes!

My short tour of these tools and instruments (and of course there are many others that could be mentioned) is meant to point up both possibilities and limitations. We are better off with the CLA, the NSSE, and other new tools than without them. But the bottom line is that the instruments now available for accountability purposes are necessarily short, superficial, and limited. They are designed to interfere minimally with instruction and to be sufficiently general and unrelated enough to the details of any institution's curriculum that they can be broadly used. In vivid contrast, the great promise of assessment is its deployment in the service of instruction, its capacity to inform the judgment of faculty and students regarding how they can best advance the quality of learning. So the challenge before us is to develop systems of assessment and accountability in which the internal uses of assessment for instruction—and the external uses of assessment for accountability and transparency—are carefully weighed. Ultimately, these are the books that need to be balanced—or, when necessary, to be strategically unbalanced.

So what are the lessons to be learned from our sense of accountability as narrative and argument? What tools and approaches can provide the most valid account of the condition of higher education and its constituent institutions? Is the most valid account necessarily the broadest and most comprehensive? Is the best strategy to develop highly specific, narrowly targeted instruments that offer deep insight into particular kinds of learning and development? Should we be looking at institutional performance or at the learning of individual students?

Seven Pillars of Assessment for Accountability

Most of the principles I want to offer here are familiar, even venerable. The fact that they remain pertinent suggests how persistent many of the challenges of assessment remain.

1. Become explicit about the story you need to tell and the rationale for choosing it. An account is one story among the many that could be told about the quality and character of an educational experience. No instrument can claim validity, no account can earn a warrant, without a clear explanation of why this story is being told instead of others. Indeed, it should be clear what the major alternative accounts could be and why they were rejected. Any one form of assessment, however rich, is a compromise, a choice among a set of legitimate possibilities.

2. Do not think that there is a "bottom line." An early step in the deployment of any instrument, new or old, should be a process of locating the instrument in a larger conceptual framework that explicitly stipulates what it does measure and what it does not. Since there is no real bottom line, the first obligation of the person rendering an account is to take responsibility for locating its unavoidable insufficiencies. Shavelson does this quite clearly for the CLA in this issue of Change, locating its domains of measurement within a figure that sketches out the broader domains that it does not assess. Bloom's classic taxonomies provide tools that can be employed in a similar manner.

Moreover, judgments of validity are never a property of measuring instruments per se. Validity can only be judged when we examine assessment results in the context of a particular argument or narrative. The cardinal principle of accountability is that counting is only meaningful and useful in the context of valid recounting. Indeed, we might make a distinction between measurement and assessment in this regard, with assessment referring to the manner in which one arrays, displays, and interprets particular measurements in the service of judgments, decisions, and actions.

3. Design multiple measures. As the stakes associated with a measurement rise, the restrictions on its form rise concomitantly—thus the need to move from judgment to measurement and from interpretation to objectivity. But as in any form of social inquiry, the price of precision is narrowness of scope. Therefore, a third principle that follows from the "no-bottom-line" observation is that nearly any use of assessment for serious practical and policy guidance should intentionally employ an array of instruments that will constitute a "union of insufficiencies." It is dangerous to permit highly consequential decisions of policy and practice to rest on the results of a single instrument, however carefully it has been field-tested and ostensibly validated.

In the Texas system of accountability for colleges and universities, for example, more than a dozen instruments are recommended for use, including the National Survey of Student Engagement, the Collegiate Learning Assessment, and multiple indices of access, graduation, and post-graduation success, often broken down by the racial and ethnic backgrounds of the students. Using this array of indicators enables others to render accounts that respond to their questions.

4. Work on combining multiple measures. A fourth principle is that a set of instruments, each with is own scores, indices, and observations, will deliver on its promise only if we take on the hard task of developing rules for deciding how to display, organize, and aggregate those indicators for making decisions. Inevitably, those decisions are functions of human judgment—which is, after all, an essential element in any such process, not something to be feared or avoided. On the other hand, there is a good argument to be made for "mechanical combination," in which general policies are debated and determined in such a way that algorithms for systematically combining the available data can be computed objectively. The late Hillel Einhorn of the University of Chicago referred to this process as "expert measurement and mechanical combination."

5. Remember that high stakes corrupt. A fifth principle is that high stakes attached to assessments have a tendency to distort the educational and evaluation processes they were intended to support. This is not only because teachers and students are sorely tempted to cheat when the stakes are high. It is also because when test designers know that high stakes are involved, they have a tendency to use items less likely to be uncertain or subject to competing judgments and arguments. As the instruments are weeded of such items or sections, they gain reliability and objectivity but often at the sacrifice of validity and nuance.

The most significant feature of high-stakes assessment is this: The higher the stakes, the greater the likelihood that teachers will teach to the test. These assessments must be designed so that the tests are worth teaching to. This is not a trivial challenge. It cries out for a strategy of embeddedness.

6. Embed assessment into ongoing instruction. Assess early and assess often. In my early days in Chicago, we used to joke, "Vote early and vote often." High-stakes assessments are likely to be used very late in the course or program where they are employed in the service of accountability. But the later the assessment, the later the knowledge of results, and the less likely it is that the assessments will yield information that can guide instruction and learning. I call these "high-stakes/low-yield" forms of assessment. They may satisfy accountability mavens but have little educative value. Instead, we should develop low-stakes/high-yield forms of assessment, much like the "running records" used by K-12 reading teachers or the routine medical history, physical examinations, or lab tests that physicians and nurses administer.

Assessment should not only serve as an external evaluation and public conscience for higher-education institutions; at the very least, it also should do no harm to instruction, and at best, it should guide, support, and enrich it. When we embed assessment in instruction, it is much more likely that what is assessed will contribute to and be compatible with the core objectives of instruction. If colleges and universities can become active pedagogical laboratories, assessment that is useful for both instruction and accounting will be actively embedded and used continuously.

Embedded measures will necessarily be designed with a different "grain size" from those designed exclusively for external, high-stakes assessments. They will be more particular than general; more dedicated to measuring individual student progress than institutional success; repeatedly administered rather than being single end-of-course events; and highly transparent to students and teachers. They will have quick turn-around times rather than providing the highly secure, secretive, and delayed feedback of current high-stakes environments. This is assessment as a regular physical exam rather than as a public autopsy.

This aspect of assessment emphasizes the need for bilateral transparency. That is, the progress students are making needs to be as accessible to them as it is to teachers or policymakers. Such transparency can empower students to take greater control of their own destinies. It is, after all, ultimately the student who must own her or his understanding and progress. Systems of assessment that are opaque, secretive, and slow-responding cripple students' sense of responsibility.

7. Become an active and collaborative site for research on new forms of assessment, new technologies to support such work, and better strategies for integration of such approaches with instruction. If the use of single-instrument, high-stakes/low-yield assessment tools will, as many of us have argued over the years, undermine the most important goals and purposes of education, then those of us who design and deploy assessments have a professional and ethical responsibility to design them to contribute more positively to the quality of teaching and learning for all students. The need now is for new assessment research and development, a project that can succeed only if institutions collaborate, experiment, and open their windows so that national work can move our fields ahead.

We need a strategy to combine the local with the national and to meld low-stakes assessment with an accountability approach that will be minimally corrupting. This will require a change in the reward system of higher education to encourage faculty to engage in such experimental approaches to their teaching, rather than worrying that they will be punished if they permit such activity to interfere with more traditional forms of research and scholarship. In the public-policy arena, the culture of competition and ranking, of punitive reactions to honest accounting, of oversimplification via report cards and bottom lines must be resisted.

Taking Control of the Narrative

One of the reasons Dina was so taken with the metaphor of narrative in accounting was that the career she had pursued just before her MBA program was as a psychotherapist. During her graduate study in social work, she had been drawn to "narrative therapy" as an approach to counseling. In narrative therapy, the central idea is that each one of us is living the life of a character in a play or a novel. Some of us feel that we have a great deal of influence over the plot of the play, while others, alas, feel that they are characters in someone else's drama. The goal of the psychotherapy is to support one's clients in seeing the narratives they feel they are living but have no control over, and to develop strategies for becoming the authors of their own stories, able to act responsibly in the situation and exercise real agency over their lives.

I often feel that academics, in the face of the growing volume of calls for accountability, have developed a sense of higher education as victim, swept away by a powerful current over which we can exercise little influence. We think of accountability as a sinister plot invented by others, controlled by the enemy, and designed to take over our professional lives and make us unhappy. We must either paddle upstream, resisting all the way, or just go with the flow, adopting a stance of minimal compliance while hoping to find a little eddy in which we can float about undisturbed. But skilled white-water rafters and canoeists remind us that neither paddling against the current nor going with the flow is a particularly fruitful tactic. The best way to get where you want to go when negotiating the rapids in a fast-moving stream is to paddle faster than the current.

In this spirit, our responsibility is to take control of the narrative. We educators must take advantage of the deep connections between counting and recounting to define the characters, the plots, the foreground, and the background of the new accountability systems. We must summon the creative energy and ambition to take advantage of the momentum (and resources)

unleashed by the new policies and exploit them to initiate the long-overdue progress in assessment needed to improve the quality of learning in higher education.

We are obligated to recount the narratives of most interest to our key stakeholders, but we cannot be limited to those alone. We must display the evidence of teaching and learning (and their embarrassments) through the multiple legitimate narratives we create about our work and our students' fates. We must account for higher-order understanding and critical thinking, in addition to factual knowledge and simple skills. We must tell of the development of civic responsibility and moral courage, even when our stakeholders have not thought to ask for those books.

Moreover, we must make the process through which we render the accounts transparent to our stakeholders. The most important of these stakeholders are our students, who need to feel a sense of agency and responsibility in this relationship as well.

The current quest for accountability creates a precious opportunity for educators to tell the full range of stories about learning and teaching. Counting and recounting can only be pursued together. Counting without narrative is meaningless. Narrative without counting is suspicious. We now have an opportunity to employ the many indicators of learning that we can count for the most important stories we have to tell.

Resources

Hillel Einhorn, *Organizational Behavior and Human Performance 7* (1972): 86-106.

Mary Poovey, *A History of Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. Chicago: University of Chicago Press, 1998.