

Principles for the Uses of Assessment in Policy and Practice
President's Report to the Board of Trustees of the Carnegie Foundation for the
Advancement of Teaching

Lee Shulman

One of the most dangerous ideas in assessment is the myth of a “magic bullet,” some powerful test with psychometric properties so outstanding that we can base high-stakes decisions on the results of performance on that measure alone. Ironically, the very opposite is true. The better the arguments we can make regarding the validity of any given measure, whether of achievement, aptitude, and any other virtue, the less appropriate it is as the sole basis for consequential decisions.

It is dangerous to permit highly consequential decisions of policy and practice to rest on the results of any single instrument, however carefully it has been field tested and ostensibly validated. I say “ostensibly” because validity brings with it, ironically, a necessary insufficiency. That is, to achieve validity, test designers have to narrow the focus of any instrument to a sobering degree. Therefore, 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 measures—and *what it does not*. Shavelson and Huang, in their paper for the ETS/CFAT conference, do this quite clearly for the Collegiate Learning Assessment (CLA)—an instrument designed to measure complex critical thinking outcomes of a liberal education and has been under development and pilot use for the last several years. They locate the CLA's domains of measurement within a framework that explicitly calls our attention to what it does not assess, as well as what it directly measures. Bloom's classic taxonomies of educational objectives provide tools to employ in a similar manner. My own Table of Learning can also be used for the purpose of defining what any given assessment does and does not measure. So the first lesson regarding an assessment is to take responsibility for locating its unavoidable insufficiencies in relation to what it claims it can measure.

The second principle, which follows from the first, is that nearly any goal of using the results of assessment for serious practical and policy guidance should intentionally employ an array of instruments that will constitute a “union of insufficiencies.” Thus, in the Texas system of accountability for higher education institutions, 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, post-graduation success, etc. This is a wise, indeed a mandatory strategy. We do not seek one perfect measurement instrument, but an array of indicators that can be understood in relation to one another.

A third principle is that an array of instruments, each with its own scores, indices and observations, will be fully useful only if we can develop appropriate decision rules for displaying, organizing and aggregating these indicators for making decisions and judgments. On the one hand, it is important to remember that these decisions and judgments are just that: exercises in human judgment; we should not back away from the

need for judgment. Indeed, economists have for many years employed a set of leading economic indicators to evaluate the health of the economy. Similarly, physicians make judgments about our physical health by gathering data on a host of items of personal history, physical examination and laboratory tests. They then exercise professional judgment in integrating these into an assessment of the state of our health.

On the other hand, our judgments can be arguably be strengthened through a process of “mechanical combination” in which general policies are debated and determined, and then combinatorial rules for using the available data can be computed objectively. We humans may not be the best combiners of data in complex settings; we can be easily fooled by our intuitions. The late Amos Tversky and his colleague Daniel Kahneman conducted pioneering research which led to the latter’s receipt of the Nobel Prize in Economics by demonstrating the fallibility of our intuitions in such circumstances. The late Hillel Einhorn of the University of Chicago referred to this hybrid strategy as “Expert Measurement and Mechanical Combination.”

A fourth principle is that high stakes attached to assessments have a tendency to corrupt 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. When test designers know that high stakes are involved they have a tendency to use items less likely to be uncertain and subject to competing judgments and arguments. As the instruments are weeded of such items or sections, they gain reliability and objectivity, but often at a sacrifice of validity.

Associated with this principle is the likelihood that high stakes assessments are too often deployed only very late in the period of educational opportunity for which accountability is exercised. 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 the instruction and the learning. I call these **high stakes/low yield** forms of assessment, and they may satisfy accountability mavens but fall far short of meeting the standards of educative value. Instead, we should seek the development of **low stakes/high yield** forms of assessment, much like the running records used by reading teachers or the physical exam or lab tests used by physicians and nurses.

The goal of low stakes/high yield assessments that are embedded in the flow of instruction in an integral manner demands existence proofs since at the moment it is somewhat counter-intuitive and contrary to standard practice. But we can cite at least two examples from the work of members of the Carnegie board.

Pat Cross has pioneered in leading the movement toward “classroom research” in higher education. The assessment strategies and tools she advocates, such as the legendary “one-minute paper,” are designed to be ongoing, embedded, formative and low stakes. A question that has not been asked of that work, as far as I know, is whether one could examine a full semester’s compilation of one minute papers and make strong inferences about the quality of learning, among individuals and across the class—though in a sense,

of course, this is exactly what we are doing with the scholarship of teaching and learning, which very much builds on Pat's work.

In the In2Books program, Nina Zolt and her colleagues have students writing repeated letters to pen pals about the books they are reading jointly across the year. No single letter has any "stakes" attached to it. Each is evaluated using a common, well-defined and validated analytic rubric, and is therefore scored on a number of dimensions. I would expect that the aggregate performance of elementary school students in the In2Books program across five different writing samples is a far more valid indicator of their growth in literacy than a single, high-stakes administration of a standardized achievement test at the end of the year. Ironically and perversely, policy makers require that programs like In2Books, with exemplary embedded and repeated low stakes/high yield assessments, are required to validate their claims against an external, one-shot, high stakes assessment.

Finally, if the use of single-instrument, high stakes/low yield assessment tools is so predictably insufficient and potentially corrupting in its consequences, those of us who design and deploy assessments have an ethical and moral responsibility to design them as elements of a suite of assessments so they contribute more positively to the quality of teaching and learning for all students.