By Stephen Klein and Richard Shavelson and Roger Benjamin

As participants in the debate regarding appropriate strategies for assessing learning in higher education, we agree with some of the statements Trudy Banta made in her Inside Higher Ed op-ed: “A Warning on Measuring Learning Outcomes.” For example, she says that “it is imperative that those of us concerned about assessment in higher education identify standardized methods of assessing student learning that permit institutional comparisons.” We agree. Where we part company is on how that can best be achieved.

Banta recommends two strategies, namely electronic portfolios and measures based in academic disciplines. One of the many problems with the portfolio strategy is that it is anything but standardized and therefore unable to support institutional comparisons. For instance, the specific items in a student’s portfolio and the conditions under which those items were created (including the amount and types of assistance the student received) will no doubt differ across students within and between colleges. In short, the portfolio is not standardized and therefore cannot function as a benchmark for institutional comparisons.

The problem with Banta’s second strategy, discipline specific measures, stems from the vast number of academic majors for which such measures would have to be created, calibrated to each other (so results can be combined across majors), and updated, as well as the wide differences of opinion within and between institutions as to what should be assessed in each academic discipline. Banta is concerned that “if an institution’s ranking is at stake [as a result of its test scores], faculty may narrow the curriculum to focus on test content.” However, that problem is certainly more likely to arise with discipline specific
measures than it is with the types of tests that she says should not be used, such as the critical thinking and writing exams employed in the Collegiate Learning Assessment (CLA) program, with which we are affiliated.

Thus, while we agree with Banta that there is a place for discipline specific measures in an overall higher education assessment program, the CLA program continues to focus most of its efforts on the broad competencies that are mentioned in college and university mission statements. These abilities cut across academic disciplines and, unlike the general education exams Banta mentions, the CLA — which she does not mention by name, but is implicitly criticizing — assesses these competencies with realistic open-ended measures that present students with tasks that all college graduates should be able to perform, such as marshalling evidence from different sources to support a recommendation or thesis (see Figure 1 for sample CLA scoring criteria and this page for details).

We suspect that Banta’s criticism of the types of measures used in the CLA program stems from a number of misperceptions about their true characteristics. For example, Banta apparently believes that scores on tests of broad competencies would behave like SAT scores simply because they are moderately correlated with each other. However, the abilities measured by the CLA are quite different from those assessed by the general education tests discussed in Banta’s article, such as the SAT, ACT and the MAPP. Consequently, an SAT prep course would not help a student on the CLA and instruction aimed at improving CLA scores is unlikely to have much impact on SAT or ACT scores.

Moreover, empirical analyses with thousands of students show that the CLA’s measures are sensitive to the effects of instruction; e.g., even after holding SAT scores constant, seniors tend to earn significantly higher CLA scores than freshmen. Differences are in the order of 1.0 to 1.5 standard deviation units. These very large effect sizes demonstrate that the CLA is not simply assessing general intellectual ability.

Banta also is concerned about score reporting methods, such as those used by the CLA, that adjust for differences among schools in the entering abilities of their students. In our view, score reporting methods that do not make this adjustment face very difficult (if not insurmountable) interpretative challenges. For example, without an adjustment for input, it would not be feasible to inform schools about whether their students are generally doing better, worse, or about the same as would be expected given their entering abilities nor whether the amount of improvement between the freshmen and senior years was more, less or about the same as would be expected.

The expected values for these analyses are based on the school’s mean SAT (or ACT) score and the relationship between mean SAT and CLA scores among all of the participating schools. This type of “value added” score reporting focuses on the school’s contribution to improving student learning by controlling for the large differences among colleges in the average ability of their entering students.

Banta objects to adjusting for input. She says that “For nearly 50 years measurement scholars have warned against pursuing the blind alley of value added assessment. Our research has demonstrated yet
again that the reliability of gain scores and residual scores — the two chief methods of calculating value added — is negligible (i.e., 0.1).”

We suspect the research she is referring to is not applicable to the CLA. For example, the types of measures she employed are quite different from those used in the CLA program. Moreover, much of the research Banta refers to uses individual-level scores, whereas the CLA program uses scores that are much more reliable because they are aggregated up to the program or college level.

Nevertheless, it is certainly true that difference scores (and particularly differences between residual scores) are less reliable than are the separate scores from which the differences are computed. But how much less? Is the reliability close to the 0.1 that Banta found with her measures or something else?

It turns out that Banta’s estimates are way off the mark when it comes to the CLA. For example, analyses of CLA data reveal that when the school is the unit of analysis, the reliability of the difference between the freshmen and senior mean residual scores — which is the value added metric of prime interest — is a very healthy 0.63, and the reliability of institutional level residual scores for freshmen and seniors are 0.77 and 0.70, respectively. All of these values are conservative estimates (see Klein, et al. 2007 for details). Even so, these values are far greater than the 0.1 predicted by Banta, and they are certainly sufficient for the purpose for which CLA results are used, namely obtaining an indication of whether a college’s students (as a group) are scoring substantially (i.e., more than one standard error) higher or lower than what would be expected relative to their entering abilities.

Banta concludes her op-ed piece by saying that “standardized tests of generic intellectual skills which she defines as ‘writing, critical thinking, etc.’ do not provide valid evidence of institutional differences in the quality of education provided to students. Moreover, we see no virtue in attempting to compare institutions, since by design, they are pursuing diverse missions and thus attracting students with different interests, abilities, levels of motivation, and career aspirations.”

Some members of the academy may buy into Banta’s position that no standardized test of any stripe can be used productively to assess important higher education outcomes. However, the legislators who allocate funds to higher education, college administrators, many faculty, college bound students and their parents, the general public, and employers may have a different view. They are likely to conclude that regardless of a student’s academic major, all college graduates, when confronted with the kinds of authentic tasks the CLA program uses, should be able to do the types of things listed in Figure 1. They also are likely to want to know whether the students at a given school are generally making more or less progress in developing these abilities than are other students.

In short, they want some benchmarks to evaluate progress given the abilities of the students coming into an institution. Right now, the CLA is the best (albeit not perfect) source of that information.

Stephen Klein is director of research and Roger Benjamin president of the Council for Aid to Education, and Richard Shavelson is Margaret Jacks Professor of Education at Stanford University.
The original story and user comments can be viewed online at http://insidehighered.com/views/2007/02/08/benjamin.

© Copyright 2006 Inside Higher Ed