Listening. Learning. Leading.®

R & D Connections

No. 10 • August 2009

Measuring Learning Outcomes in Higher Education

By Ou Lydia Liu

The Voluntary System of Accountability (VSA):

- Was developed in 2007 by the American Association of State Colleges and Universities (AASCU) and the National Association of State Universities and Land-Grant Colleges (NASULGC)
- Includes, as of April 2009, 321 institutions from all 50 U.S. states
- Evaluates core educational outcomes in public colleges and universities
- Uses the term *value-added* to refer to the academic progress students make from freshman to senior year
- Uses standardized tests to measure value-added

How do we know what students have learned after they have been in college for four years or even longer?

As college tuitions and fees continue to grow, students, parents and public policymakers are interested in understanding how public universities operate and whether their investments are well utilized. Accountability in public higher education has come into focus following the attention accountability has received in K–12 education.

Under former U.S. Secretary of Education Margaret Spellings, the formation of the Commission on the Future of Higher Education highlighted the importance of accountability. The commission's report emphasized accountability as one of the four areas needing urgent attention in U.S. higher education (U.S. Department of Education, 2006).

It was against this backdrop that two leading organizations in higher education, the American Association of State Colleges and Universities (AASCU) and the National Association of State Universities and Land-Grant Colleges (NASULGC), developed a program called the Voluntary System of Accountability (VSA; **http://www.voluntarysystem.org**).

Since its inception, the VSA has received increasing attention from institutions of higher learning all across the United States. As of April 2009, 321 institutions from all 50 states had signed up for the VSA program (Voluntary System of Accountability [VSA], 2009).

The primary purpose of the VSA is to evaluate core educational outcomes in public universities and colleges by focusing on skills that are "common, multidisciplinary, and university-wide" (VSA, 2008, p. 2). The VSA defines core educational outcomes as skills in written communications, critical thinking, and analytic reasoning, and asserts that these skills are necessary for students to survive and thrive in the 21st century.

The VSA selected three standardized tests from a number of possible tests to measure these core educational outcomes. The three tests are the ETS® Proficiency Profile[1] offered by ETS, the Collegiate Assessment of Academic Proficiency (CAAP) offered by ACT®, and the Collegiate Learning Assessment (CLA) offered by the Council for Aid to Education (CAE).

These three tests were selected because they are believed to provide reliable and valid measures of critical thinking, analytic reasoning, and written communication, as broadly defined.

[1] An earlier version of this article referred to the ETS Proficiency Profile by its former name, the Measure of Academic Proficiency and Progress (MAPP). The assessment was renamed in 2009. Only the name has changed. All other aspects of the test remain the same.

Editor's note: Ou Lydia Liu is an associate research scientist in the Foundational and Validity Research area of ETS's Research & Development division.

"The assumption underlying the VSA is that universities are responsible, at least partially, for students' intellectual progress during their college years."

Value-Added in VSA

The VSA uses the term *value-added* to refer to the learning progress college and university students make from freshman to senior year. The VSA measures this by looking at the difference between freshmen and senior performance on a standardized test such as the ETS Proficiency Profile.

As part of this measurement, the VSA controls for students' admission scores on college admissions tests such as the SAT® or ACT; that is, the system accounts for the fact that some students are entering higher education already better prepared than others. The underlying assumption is that universities are responsible, at least partially, for students' intellectual progress during their college years.

There are two possible ways to measure performance differences between freshmen and seniors.

One way is to test students in their freshman year and test them again in their senior year. This way, the same group of students is tested twice. This design is referred to as a longitudinal design.

The other way is to test a group of freshmen and a group of seniors at the same time so that the freshmen and the seniors are not the same group of students. This design is referred to as a cross-sectional design.

The VSA has adopted the cross-sectional design because of its practical advantages: It is easier and less costly to test two groups of students at the same time than to track the same group of students over four years. Therefore, to participate in the VSA, an institution will administer one of the three tests to its freshmen and seniors, possibly at the same time.

To compare institutions on core educational outcomes, the VSA computes a value-added index. Researchers at CAE developed this method, which is used with the CLA test (Klein, Benjamin, Shavelson, & Bolus, 2007). Since the purpose is to evaluate institutional effectiveness, the VSA conducts the analysis at the institutional level instead of at the individual student level.

The value-added computation compares the actual learning gains at an institution from freshman to senior year with the expected learning gains given students' admission scores.

If students at an institution made a larger-than-expected learning gain, as measured by a standardized test, then this institution will be assigned a higher value-added index. Similarly, if students at an institution made a smaller-than-expected learning gain, then this institution will be assigned a lower value-added index. Essentially, this is how the institutions are compared side-by-side.

Challenges and Possible Directions for Value-Added Research

Although the value-added approach holds great promise for evaluating instructional effectiveness among public institutions, it has many inherent challenges that may affect the validity of the inferences that can be drawn from the test results.

Advancing the Value-Added Approach to Measuring Outcomes

Additional research is needed to support the validity of the inferences to be drawn from assessments of higher education effectiveness. Some important questions include:

1. How comparable are results from the three different assessments that the VSA uses to evaluate learning outcomes?
2. When selecting students who will take the standardized tests used by the VSA, how can institutions choose test takers who make up a representative sample of their student population?
3. How does test-taker motivation affect the results of tests given as part of the VSA effort?
4. Does the current method of calculating value-added provide a meaningful estimation of an institution's effectiveness?
5. How can institutions be compared fairly to each other?

For example, with the current way the three standardized tests are administered at institutions, there is no guarantee that students who take the tests are representative of that institution. Since the test results do not have a direct impact on individual students, their possible lack of motivation in taking the tests could be another concern.

Furthermore, additional research evidence is needed to support the current method of value-added calculation. A fair evaluation of public institutions requires decisions about how these issues can best be addressed.

Although unlikely to be addressed immediately, these issues should be thoroughly discussed so stakeholders understand the benefits and caveats of the current approach.

Comparability

After institutions sign up for the VSA program on a voluntary basis, they have the flexibility to choose one of the three tests as their accountability measure. Therefore, it is important to consider the comparability of results from the three tests.

Two major differences exist among the three tests: There are differences in item format (the ETS Proficiency Profile and CAAP are multiple-choice tests, while CLA is an essay-type test) and differences in delivery format (the ETS Proficiency Profile includes both a paper-and-pencil and an online version; CAAP is a paper-and-pencil test; and CLA is delivered online).

There are also other differences, such as test length and whether the test is modularized. What's more, investigations of the similarity or dissimilarity among the tests on their critical thinking and writing measures (these two skills being of key interest to the VSA) have not yet been carried out.

To understand how comparable these tests are, the organizations that develop them are undertaking a joint study to examine the construct validity of these tests — that is, they are studying whether the tests measure the same thing. The study will consider two major questions:

- What is the correlation between scores from ETS Proficiency Profile critical thinking and scores from CLA critical thinking?
- Does item format have an impact on student performance?

The study is supported by the U.S. Department of Education's Fund for the Improvement of Postsecondary Education (FIPSE).

Finding Representative Samples

To draw conclusions about an institution based on a sample of students, it is critical to ensure that this group of students represents the school's total student population in terms of race, gender, academic achievement, language, social status, and other important factors.

Institutions also use a wide variety of incentives to recruit students to take outcomes assessments. In campus advertisements designed to get test takers to sign up, colleges and institutions have offered students course credit, bookstore coupons — even a free smoothie — for their trouble.

"As an incentive to take the tests used by the VSA, institutions have offered students course credit, bookstore coupons — even a free smoothie — for their trouble."

Because students decide to take the test on a voluntary basis, there is no guarantee that they represent the institution as a whole. This raises questions about how to draw inferences from a sample to an institution. Institutional researchers should, possibly through collaboration with testing organizations, develop a mechanism to ensure sample representativeness. Otherwise, findings resulting from an unrepresentative sample should not be generalized to the entire university.

Student Motivation

Researchers are rightfully concerned about whether students will take the test seriously if the test results do not have a direct impact on them (Banta, 2008; Borden & Young, 2009). If students at a particular institution do not try their best when taking the test, the results are likely to lead to an underestimation of that institution's effectiveness.

There are some ways to monitor student effort in test taking. For example, in an online delivery format, the amount of time a student takes to answer each question can be measured. If a student is found to have consistently spent an unusually short amount of time answering items, this may be evidence that the student did not treat the test seriously. Such responses may need to be removed from analyses since they pose a threat to the validity of the results (Kong, Wise, & Bhola, 2005).

A recent study on the ETS Proficiency Profile (Liu, 2008) provides some evidence that, in general, ETS Proficiency Profile test takers display no significant variation in motivation compared with those who take a higher-stakes assessment. The correlation between mean ETS Proficiency Profile score and mean SAT score was found to be .83 on writing and .85 on critical thinking, based on data from 6,196 students at 23 institutions.

If student motivation had varied significantly in taking the ETS Proficiency Profile test, the correlation would not have been so high, since the SAT is an extremely high-stakes test and student motivation on SAT should be almost uniformly high. Although their ETS Proficiency Profile test performance does not directly affect whether or not they graduate, students may have wanted to present their institution in its best light. How their institution ranks among the competition may affect institutional reputation, which could reflect on the quality of their diploma.

Value-Added Methodology

The current value-added method includes students' admission scores as the only predictor of their performance on standardized higher education outcomes tests such as the ETS Proficiency Profile. However, there are many other factors that could influence student learning in college.

For example, students' freshmen-to-senior progress could also be affected by student gender, language status (i.e., speaking English as a first language), an institution's selectivity, or the amount of resources the institution has access to. These factors should be controlled for in the investigation of institutional effectiveness for a more meaningful estimation of student learning.

Linking Student Performance to Institutional Effectiveness

Probably one of the most important and sensitive issues in value-added research is the link between student performance and institutional effectiveness. Besides program structure and instruction at an institution, there are many other determinants of student learning, and often these factors are beyond an institution's control. For example, student motivation, family support, and financial status can all have an impact on student achievement in college.

The key question is the degree to which institutions should be held accountable for the variation in student learning that remains once other factors are considered. Therefore, we need to be very careful in linking student performance to an institution's effectiveness since a causal relationship has not yet been established.

What can we do to make this comparison fair? The answer may be to compare students in similar institutions. That is, to use the old cliché, we should compare apples to apples and oranges to oranges: It may be fair to compare less-selective Southern State University to less-selective Northern State University,[2] but not necessarily fair to compare those two against a highly selective public institution.

Summary

Despite the challenges we face, accountability is needed in higher education for the same reasons it is needed in K–12 education and in any other area of education. Because a good education has become a pathway to opportunities and success, stakeholders deserve to know whether institutions have done their best to maximize student learning and have effectively utilized public resources.

It is important to engage all stakeholders, including students, parents, faculty members, institutional administrators, testing organizations, and public policymakers, in the discussion.

These stakeholders need to reach a scientific common ground as to how institutions should be evaluated and what constituencies should be involved. This common understanding is crucial to the fruitfulness of programs such as the VSA that aim to evaluate institutional effectiveness.

All important factors that may affect student learning should be considered when we hold institutions responsible for student achievement in college. Additional research evidence is needed to identify a most accurate and meaningful way of defining and calculating value-added for the evaluation of higher education processes for accountability purposes.

References

Banta, T. (2008). Trying to clothe the emperor. *Assessment Update*, *20*(2), 3-4, 16-17.

Borden, V. M. H., & Young, J. W. (2009). Measurement validity and accountability for student learning. In V. M. H. Borden & G. R. Pike (Eds.), *Assessing and accounting for student*

[2] These generic, fictional institution names have been chosen for illustrative purposes. Any similarity to the names of real U.S. universities is coincidental.

learning: Beyond the Spellings Commission: New directions in institutional research (pp. 19-37). San Francisco: Jossey-Bass.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review*, *31*, 415-439.

Kong, X., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*, 606-619.

Liu, O. L. (2008). *Measuring learning outcomes in higher education using the Measure of Academic Proficiency and Progress (MAPP)* (ETS Research Report No. RR-08-47). Princeton, NJ: Educational Testing Service.

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. Washington, DC: Author.

Voluntary System of Accountability. (2008). *Voluntary system of accountability (VSA): Information on learning outcomes measures*. Retrieved May 1, 2009, from http://www.voluntarysystem.org/docs/cp/LearningOutcomesInfo.pdf

Voluntary System of Accountability. (2009). *VSA participants by state*. Retrieved May 1, 2009, from http://www.voluntarysystem.org/index.cfm?page=templates

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001
e-mail: **RDWeb@ets.org**

Editor: Jeff Johnson

Visit ETS Research & Development on the Web at
www.ets.org/research

Copyright © 2010 by Educational Testing Service. All rights reserved. ETS, the ETS logo and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS). All other trademarks are the property of their respective owners. 16117