



**BOARD OF TRUSTEES
& OFFICERS**

Benno Schmidt
Chairman

Roger Benjamin
President & CEO

James Hundley
Executive Vice President & COO

Richard Atkinson
President Emeritus
University of California System

Doug Bennett
President
Earlham College

Michael Crow
President
Arizona State University

Russell C. Deyo
Vice President & General Counsel
Johnson & Johnson

Richard Foster
Managing Partner
Millbrook Management Group, LLC

Ronald Gidwitz
Chairman
GCG Partners

Lewis Kaden
Vice Chairman
Citigroup Inc.

Michael Lomax
President
United Negro College Fund

Katharine Lyall
President Emeritus
University of Wisconsin System

Eduardo Marti
President
Queensborough Community College,
CUNY

Ronald Mason
President
Jackson State University

Diana Natalicio
President
University of Texas at El Paso

Charles Reed
Chancellor
California State University

Michael Rich
Executive Vice President
RAND Corporation

Farris Womack
*Executive Vice President and Chief
Financial Officer, Emeritus
Professor, Emeritus*
University of Michigan

Stephen Klein
Director of Research
CAE

October 9, 2009

A recent study supported by the Fund for the Improvement of Postsecondary Education (FIPSE) found that schools with relatively high scores on the Collegiate Learning Assessment (CLA) also had relatively high scores on the multiple choice reading, writing, and mathematics tests in the MAPP and CAAP test batteries published by ETS and ACT, respectively. As we have long argued, the choice of which tests and benchmarks to use says a great deal about the values of an institution. We are pleased that the multiple-choice tests in the FIPSE study correlated highly with the complex CLA assessments because that finding indicates the CLA school-level scores are just as reliable as those obtained with multiple-choice tests. We believe these findings will free schools to choose tests that assess and foster the kinds of thinking and learning that are consistent with the goals of higher education.

Some policymakers, however, may misinterpret these high correlations as suggesting that all of these tests measure the same knowledge, skills, and abilities. If that were so, it would only make sense to use the less expensive multiple-choice tests. However, the very high school-level correlations among all the measures does *not* mean they assess the same skills or that their scores should be combined under one banner. Instead, it simply signifies that as a group, the schools whose students are proficient in one area also tend to be proficient in the other areas tested.

There are several ways to demonstrate that different tests measure different skills even when their scores are highly correlated. This is evident from even a cursory inspection of the tests. For example, a test that requires students to solve algebra problems bears no resemblance to a CLA Performance Task, such as one in which students review charts, editorials, reports and other documents and then write a memo discussing the pros and cons of alternative plans for dealing with a community problem such as pollution in the local lake. The high correlation between the CLA and multiple-choice measures, particularly at the school level, may simply stem from the students who have the skills needed for the algebra test also having the skills needed for the CLA. The high correlation does not mean these tests measure the same skills, and it says nothing about whether a school's students are just as proficient in one area as they are in another.

Another way to illustrate that highly correlated tests may measure different things is to have a researcher ask students to "think out loud" as they go through the tests so as to document their thought processes. What emerges from this activity on a reading test is quite different than what is reported on a science test or the CLA, even though there *is* a strong correlation between the scores on all of these measures.

A third way to demonstrate that different tests measure different skills is to consider how students would be prepared to take each test. For instance, the instructional activities that would be used to train students for multiple-choice writing tests, which



tend to emphasize editing skills, are unlikely to be the same activities as those used to help students write a critique of an argument. Specifically, one of the best ways to prepare for a multiple-choice writing test is to drill conjugations and punctuation rules. In contrast, the best way to prepare to write a critique of an argument is to practice examining and composing responses to a wide range of arguments, learning about different types of logical flaws in analyses done by others, etc.

Providing all students with appropriate instruction in an area, such as math, can increase their proficiency in that area. This increase would be indicated by an improvement in their mean math test scores, and this increase could occur without any change in their mean writing test scores. Thus, if all schools improved to the same degree in math, then there would be no change in the correlation (i.e., the rank ordering of the schools) on the two tests. That is why correlations between different measures, by themselves, provide little direct information about what a test actually measures.

The kinds of tests schools and professors use signal to students and other stakeholders what the faculty consider important. In so doing, it alerts students to what they should study and learn. The question then becomes does a school's faculty want to encourage students to learn strategies for doing well on multiple-choice tests or do they want to emphasize students learning how to analyze realistic problems (like the kind they may encounter after they graduate) and communicate the results of their analyses effectively in writing? The measures schools use says a lot about their priorities and, thereby, has a significant effect on what their students are taught and what skills they develop.

The bottom line is that the FIPSE study's finding of high correlations among a variety of quite different tests does not indicate that these tests measure the same underlying skills. Moreover, how students are prepared to take one test is likely to be quite different than how they would be prepared to take another test. These considerations led the authors of the FIPSE study report to conclude that the decision about which measures a college should use should hinge on their acceptance by students, faculty, administrators, trustees, and other policy makers.



Appendix

Findings of the FIPSE Study¹

The FIPSE study examined whether the scores on different tests that were designed to measure the same skill, such as reading, correlated higher with each other than they did with the scores on tests of different skills, such as writing or math. If tests designed to measure the same skill correlate higher with each other than they do with tests of other skills, this would support interpreting the test results in the manner intended by the test publisher.

Analyses were run two ways: once when individual student scores were examined and again when school mean scores were analyzed. Student-level results are germane to making decisions about individuals, such as for remediation and academic counseling. School-level results are used to assess the effects of instructional programs, such as college-wide efforts to improve student writing or critical thinking skills.

Analyses of the student-level data generally revealed the expected pattern of correlations. That is, student scores on the MAPP and CAAP reading tests correlated higher with each other than they did with scores on tests designed to measure other abilities. However, there were some notable exceptions to the expected results. For instance, scores on the CAAP and MAPP multiple-choice writing tests correlated higher with the scores on other multiple-choice tests, such as math, than they did with the scores on tests that required students to actually write their responses. Differences in the reliability of the scores on different measures may be one important cause of these findings.

The differentiation seen in the pattern of correlations in the student-level analyses all but disappeared in the school-level analyses. This occurred, at least in part, because all the measures correlated very highly with each other at the school level. In other words, the relative standings of the schools on one test coincided very closely with their standings on all of the other tests studied.

¹ S. Klein was principal investigator for the FIPSE-funded Test Validity Study. The views expressed here are his alone. They may not reflect the positions of colleagues from ETS or ACT. The findings summarized here relate to the discussion above only. The study findings are documented in: Klein, S., Liu, O.L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., Nemeth, A., Robbins, S., & Steedle, J. (2009). *Test Validity Study Report*. Supported by the Fund for the Improvement of Postsecondary Education. Available: www.voluntarysystem.org.

