

Test Validity Study (TVS) Report

Executive Summary

Supported by the Fund for the Improvement of Postsecondary Education
(FIPSE)

September 29, 2009

Primary Authors

Klein, Stephen (CAE)

Liu, Ou Lydia (ETS)

Sconing, James (ACT)

Secondary Authors

Bolus, Roger (CAE)

Bridgeman, Brent (ETS)

Kugelmass, Heather (CAE)

Nemeth, Alexander (CAE)

Robbins, Steven (ACT)

Steedle, Jeffrey (CAE)

EXECUTIVE SUMMARY

Purpose

This study examined whether commonly used measures of college-level general educational outcomes provide comparable information about student learning. Specifically, do the students and schools earning high scores on one such test also tend to earn high scores on other tests designed to assess the same or different skills? And, are the strengths of these relationships related to the particular tests used, the skills (or “constructs”) these tests are designed to measure (e.g., critical thinking, mathematics, or writing), the format they use to assess these skills (multiple-choice or constructed-response), or the tests’ publishers?

We also investigated whether the difference in mean scores between freshmen and seniors was larger on some tests than on others. Finally, we estimated the reliability of the school mean scores on each measure to assess the confidence that can be placed in the test results. We anticipate our findings will be useful to policy makers when interpreting test results and deciding which test(s) to use. We also expect that our findings will be of interest to test publishers and those involved in evaluating institutions and programs.

Procedures

We administered 13 different tests. These tests were among those in the ACT’s Collegiate Assessment of Academic Proficiency (CAAP), the Council for Aid to Education’s Collegiate Learning Assessment (CLA), and the Educational Testing Service’s Measure of Academic Proficiency and Progress (MAPP). These are among the most widely used college-level tests of general educational skills. Four of the tests were in critical thinking, two in reading, two in mathematics, four in writing, and one in science. Nine tests used a multiple-choice format and four used a constructed-response (open-ended or essay) format.

All 13 tests were administered at each of the study’s 13 schools. Over 1,100 students (freshmen and seniors) participated in the research. The schools varied in size, average college admission test scores, geographic region, control (public or private), and other characteristics.

All three testing agencies (ACT, CAE, and ETS) worked collaboratively and collegially in designing the study, analyzing the data, and interpreting the results.

We conducted some analyses using student-level data because test results are often considered in making decisions about individuals, such as identifying areas where they need remediation or whether they are ready to move on to more challenging courses. We also conducted analyses at the school level because test results are more reliable at that level and may be used to inform policy, resource allocation, and programmatic decisions, such as by indicating whether the progress students are making at a college is commensurate with

the progress of students at other colleges and universities or whether the progress within a school in one area (such as writing) is greater than it is in other areas.

Research Questions and Major Findings

The three questions we studied and the answers to them are discussed below.

1) What are the relationships among scores on commonly used college-level tests of general educational outcomes? Are those relationships a function of the specific skills the tests presumably measure, the tests' formats (multiple-choice or constructed-response), or the tests' publishers?

A high positive correlation between two tests indicates that students (or schools) that obtain high scores on one test also tend to obtain high scores on the other test. We found that the pattern of student-level correlations generally supported the measures' construct validity. That is, two tests of the same construct (such as reading) usually (but not consistently) correlated higher with each other than they did with measures of different constructs provided the response format (multiple-choice or constructed-response) was taken into consideration.

There was far less evidence of construct differentiation when the school was the unit of analysis. The mean school-level correlation among the nine multiple-choice tests was 0.92, which is similar to the 0.84 mean correlation among the four constructed-response measures. The latter correlation is also nearly identical to the mean school-level correlation of 0.85 between multiple-choice tests of one construct and constructed-response tests of other constructs. Taken together, these results suggest that when the analysis is conducted at the school level, all the tests order schools similarly, regardless of which constructs they are designed to measure or which response format is used.

The 0.08 difference in the average correlation between the multiple-choice and constructed-response tests may be attributable to the higher reliability of the multiple-choice scores but also to the uniqueness of the constructed-response measures. In other words, the skills required to do well on different multiple-choice tests may be more alike than the skills required to do well on different constructed-response tests. For example, the CLA's Performance Task requires the examinee to view a "document library," whereas the CLA's Critique-an-Argument Task does not. If such differences are important (and the test publisher and others think they are), then a school's relative standing could be influenced by which constructed-response measure(s) it uses.

2) Is the difference in average scores between freshmen and seniors related to the construct tested, response format, or the test’s publisher?

Answering this question involved creating an index (called “effect size”) that allowed us to compare score gains between freshmen and seniors in a way that controlled for differences in score distributions among tests as well as any differences in average SAT and ACT scores between freshmen and seniors. Larger effect sizes indicate greater differences in mean scores between classes.

Seniors had higher mean scores than freshmen on all the tests except the CAAP Mathematics test. (Effect sizes express mean differences in standard deviation units.) When this test was excluded from the analysis, effect sizes ranged from about one quarter to one half of a standard deviation. Effect sizes were not systematically related to the constructs tested, response format, or test publisher. For example, the average effect size across constructs for the ACT, CAE, and ETS measures were 0.33 (excluding mathematics), 0.31, and 0.34, respectively (see report Table 4b for details).

3) What are the reliabilities of school-level scores on different tests of general education learning outcomes?

School-level reliability refers to score consistency (i.e., a school receiving a similar mean score regardless of the sample of students taking the test). Reliability is reported on a scale from 0.00 to 1.00, where higher values indicate greater reliability.

With schools as the unit of analysis, score reliability was high on all 13 tests (mean was 0.87 and the lowest value was 0.75). Thus, score reliability is not a major concern when using school level results with sample sizes comparable to those obtained for this study.

Conclusions

Overall, when the school was the unit of analysis, there were very high correlations among all the measures, very high score reliabilities, and consistent effect sizes. These findings held across the test constructs measured, response formats, and test publishers. For instance, the correlation between two multiple-choice reading tests was essentially the same as their correlations with other multiple-choice and constructed-response tests of the same or other constructs. When the student was the unit of analysis, correlations among measures and reliabilities were generally but not always high; and, as expected, lower than they were when the school was the unit of analysis.

The very high correlations among all the tests at the school level could be due to different tests assessing overlapping or interrelated skills or from one skill set being dependent on another set. For example,

good writing requires critical thinking skills. The high correlations among the measures at both the school and student levels also could stem from the fact that many students who have the abilities needed to achieve in one area also have the skills necessary for other areas.

The correlations between different tests are affected by the reliability of their scores. This is not a concern at the school level because all the reliabilities at that level are quite high. However, when the individual student is the unit of analysis, multiple-choice measures are known to yield more reliable scores per hour of testing time than do constructed-response measures. One implication of these findings is that when scores are used to make decisions about individual students, such as for course placement, special attention should be given to their reliability. Similarly, drawing conclusions about a student's relative strengths across skill areas (whether measured by multiple-choice or constructed-response tests) should be limited to instances where the differences are statistically significant.

Finally, given the findings above and particularly the high correlations among the measures at the school level, the decision about which measures to use will probably hinge on their acceptance by students, faculty, administrators, trustees, and other policy makers. There also may be trade-offs in costs, ease of administration, breadth of constructs measured, and the utility of the different tests for other purposes, such as to support other campus activities and services. Indeed, the testing program may include guidance on the interpretation of results and their implications for programs and activities that complement the testing program's goal of improving teaching and learning.