Test Validity Study (TVS) Report

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)

CONTENTS

EXECUTIVE SUMMARY……2

INTRODUCTION……6

RESEARCH QUESTIONS……9

MEASURES……11

ADMINISTRATIVE METHODS……16

STATISTICAL METHODS……20

RESULTS……23

CONCLUSIONS……30

APPENDICES……35

A: SAMPLE ITEMS

B: SCHOOL AND TVS SAMPLE CHARACTERISTICS

C: SUMMARY OF FILE LINKAGE PROCEDURES

D: TVS TESTING COUNTS BY TEST AND CLASS

E: DIFFERENCES IN FRESHMAN TO SENIOR CORRELATIONS

F: EQUATIONS REFERENCED IN STATISTICAL METHODS AND RESULTS

G: LOWEST AND HIGHEST AVAILABLE CORRELATIONS

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.

INTRODUCTION

Papers and articles extolling the benefits of a college education are not hard to find. Students are told early and often that getting a college education is key to success. U.S. Census figures show that getting a bachelor's degree is associated with an increase in median income of 67% over a high school graduate (U.S. Census Current Population Survey, 2008). This, along with studies showing the relationship between education and longevity, health, and happiness, provides ample evidence that there are rewards associated with getting a college degree.

It is also true that college is expensive. A recent report by the College Board indicates that the average cost of tuition, fees, room and board for four years at a public college exceeds $55,000, while the average cost at a private institution is more than double that figure (Trends in College Pricing, 2008). Given their major investments in higher education, students, parents, and politicians want assurance that they are getting substantial educational returns in exchange for their time and money.

Concurrently, colleges have an interest in measuring academic outcomes. College faculty and administrators are committed to providing high quality education and many are interested in knowing whether their students are acquiring the knowledge and skills expected of a college graduate. Assessment results can help inform efforts aimed at improving teaching and learning by providing formative feedback about the strengths and weaknesses of academic programs.

In addition, institutional assessment programs provide schools with the opportunity to demonstrate to the public that they are successful in their educational missions. In order to assist colleges in such efforts, the Association of Public and Land-grant Universities (APLU; formerly known as the National Association of State Universities and Land-Grant Colleges) and the American Association of State Colleges and Universities (AASCU) developed the College Portrait of Undergraduate Education, which is one part of the Voluntary System of Accountability (VSA: http://www.voluntarysystem.org/index.cfm). The College Portrait is a web-based system that uses a common template and definitions to provide the public with a standardized view of information such as institutional characteristics and the results of college outcomes testing for participating institutions.

The College Portrait includes a Student Learning Outcomes component, which reports on gains in critical thinking and writing skills as students advance from freshman to senior year. To obtain this information, students are tested as they enter and as they leave an institution. Differences in performance between freshmen and seniors can reasonably be attributed, in large part, to the experience of attending that institution. By controlling for the entering academic ability (e.g., admission test scores) of both freshman and seniors who test, a "value-added score" may be computed that allows for cross-institutional comparisons of the freshman-senior difference.

Schools have three options when selecting an assessment for the Student Learning Outcomes portion of College Portrait: the Collegiate Assessment of Academic Proficiency (CAAP), published by ACT; the Collegiate Learning Assessment (CLA), published by the Council for Aid to Education (CAE); and the Measure of Academic Proficiency and Progress (MAPP), published by the Educational Testing Service (ETS). These assessments differ in format and content. When constructing an assessment, each publisher specifies the test requirements in different ways; this is based on decisions regarding the relative advantages and disadvantages of different approaches. For example, a test can be multiple-choice or constructed-response. The multiple-choice option allows for more questions to be asked in the same time period, and so these types of assessments tend to be more reliable (i.e., providing consistent scores across repeated test administrations). Multiple-choice questions are also easy and inexpensive to machine score, making this a more efficient option in certain respects. On the other hand, constructed-response items allow for greater flexibility in the type of questions asked and offer students different opportunities to demonstrate their skills. With constructed-response items, it is possible to ask questions that require students to identify a problem and formulate solutions or to pose problems that require multiple steps to solve. This approach is said to have greater "face validity" because constructed-response tasks may resemble the sorts of problems that students will deal with in the world of work and their every-day lives after graduating. It is also possible to give a student credit for solving part of a problem, which is typically not possible with a multiple-choice question.

Colleges choose a test using several criteria: alignment of the test with the educational goals of the institution, the technical qualities of the test, the cost and ease of administration, etc. Given that the VSA considers it desirable to compare performance across colleges, the fact that different institutions use different assessments may be problematic. Comparing student learning outcomes across institutions does not seem appropriate if the basis for the comparison differs across schools. As an example, consider different writing tests included in this study. MAPP and CAAP have multiple-choice writing tests, but the CLA and CAAP include writing production components.[1] Each test taps different aspects of writing skills, so it may not be appropriate to compare scores across tests.

The Test Validity Study (TVS) set out to address this challenge (and others) concerning the interpretation of scores from three tests of college learning outcomes: CAAP, CLA, and MAPP. Data collection for this study involved administering each test to freshmen and seniors from a sample of colleges. Analyses were carried out to examine the pattern of correlations among the tests, the effects of construct and response format on the strength of those correlations, the sensitivity of the tests to freshman-senior differences, and the reliability of the tests. In summary, the TVS set out to:

[1] Although MAPP also has an optional essay, it is structurally fairly similar to the CAAP Essay, and so was not included in the current study to preserve testing time.

- Gather evidence to support the valid interpretation of scores produced by measures of collegiate learning. This information would contribute to the usefulness, acceptance, and sustainability of these assessments.
- Determine the extent to which different measures of the same construct of collegiate learning correlate with each other.
- Use results from different measures of collegiate learning to investigate the utility of these measures as comparable benchmarks for monitoring student learning and informing the improvement of teaching and learning.

RESEARCH QUESTIONS

This Test Validity Study addressed three primary research questions:

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

A "construct" is a type of knowledge, skill or ability that a test is designed to measure. This study relied on multiple-choice tests to assess three constructs (reading, mathematics, and science). Two other constructs (critical thinking and writing) were measured with both multiple-choice and constructed-response tests. This first research question was addressed by examining patterns of correlations among the TVS measures to look for evidence that correlations between tests of the same or similar constructs are higher than correlations between tests of different constructs. For example, the correlation between two critical thinking tests should be relatively high compared to the correlation between a critical thinking test and a mathematics test.

Second, the effect of item response format on the correlations was examined. MAPP and CAAP tests are multiple-choice tests (except for CAAP Writing Essay), and the CLA employs a constructed-response format. One might expect correlations between tests to be higher if they share a common response format. Correlations were examined to determine whether tests measuring similar constructs and using the same response format have higher correlations than tests measuring similar constructs but using different response formats. Similar comparisons were made for tests measuring different constructs. Lastly, comparisons were made across the measures of different test publishers.

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

One of the most important goals of outcomes assessment is to provide information on student progress in college. For this study, an effect size reflecting the performance difference between freshmen and seniors, after controlling for prior academic achievement, was used as an indicator of student learning gains in college. These effect sizes were computed for each test and compared to determine if they are comparable across tests that measure the same or similar constructs, as well between tests that measure different constructs.

3. What are the reliabilities of school-level scores on different tests of college learning?

School-level reliabilities reflect the consistency of a school's mean score across theoretical repeated examinations with different samples of students. Split-half methods were used to examine score reliability at the school level. This approach involved randomly splitting students at the same institution into two groups and computing a mean for each group. The two means from each institution were then correlated across all of the institutions in the sample. The random splitting of students into two groups was replicated 1,000 times to obtain the expected value of school-level reliability.

MEASURES

Introduction

Three assessments of collegiate learning were administered as part of the Test Validity Study: CAAP, CLA, and MAPP. Each assessment is comprised of several tests. A total of 13 tests (across all three assessments) measuring five different constructs were administered as part of the Test Validity Study. Although each test is classified as within only one construct, it is recognized that a single test may measure multiple constructs and that constructs may overlap. Descriptions of each test appear below. Please refer to *Appendix A: Sample Items* for example questions from each test.

CAAP

ACT's Collegiate Assessment of Academic Proficiency (CAAP) provides scores reflecting students' skills related to critical thinking, science, reading, writing skills, essay writing, and mathematics. With the exception of the essay writing test, all of these measures use a multiple-choice format.

CAAP Critical Thinking. The CAAP Critical Thinking Test is a 32-item, 40-minute test that measures students' skills in clarifying, analyzing, evaluating, and extending arguments. An argument is defined as a sequence of statements that includes a claim that one of the statements, the conclusion, follows from the other statements. The Critical Thinking Test consists of four passages that are representative of the kinds of issues commonly encountered in a postsecondary curriculum. A passage typically presents a series of subarguments in support of a more general conclusion or conclusions. Each passage presents one or more arguments using a variety of formats, including case studies, debates, dialogues, overlapping positions, statistical arguments, experimental results, or editorials.

CAAP Science. The CAAP Science Test is a 45-item, 40-minute test designed to measure students' skills in scientific reasoning. The contents of the Science Test are drawn from biological sciences (e.g., biology, botany, and zoology), chemistry, physics, and the physical sciences (e.g., geology, astronomy, and meteorology). The test emphasizes scientific reasoning skills rather than recall of scientific content or a high level of skill in mathematics or reading. The test consists of eight passage sets, each of which contains scientific information and a set of multiple-choice test questions. A passage may conform to one of the three different formats: data representation, research summaries, or conflicting viewpoints.

CAAP Reading. The CAAP Reading Test is a 36-item, 40-minute test that measures reading comprehension as a combination of skills that can be conceptualized in two broad categories: Referring Skills and Reasoning Skills. The Reading Test consists of four prose passages of about 900 words each that are representative of the level and kinds of writing commonly encountered in college curricula. The four reading

passages come from the following four content areas, one passage from each area: Prose Fiction—Entire stories or excerpts from short stories or novels; Humanities—Art, music, philosophy, theater, architecture, or dance; Social Studies—History, political science, economics, anthropology, psychology, or sociology; and Natural Sciences—Biology, chemistry, physics, or the physical sciences.

CAAP Writing Skills. The CAAP Writing Skills Test is a 72-item, 40-minute test measuring students' understanding of the conventions of standard written English in punctuation, grammar, sentence structure, strategy, organization, and style. Spelling, vocabulary, and rote recall of rules of grammar are not tested. The test consists of six prose passages, each of which is accompanied by a set of 12 multiple-choice test items. A range of passage types is used to provide a variety of rhetorical situations.

CAAP Essay Writing. The CAAP Writing Essay Test is predicated on the assumption that the skills most commonly taught in college-level writing courses and required in upper-division college courses across the curriculum include: Formulating an assertion about a given issue; Supporting that assertion with evidence appropriate to the issue, position taken, and a given audience; Organizing and connecting major ideas; and Expressing those ideas in clear, effective language. The model developed by ACT for the Writing Essay Test is designed to elicit responses that demonstrate a student's ability to perform these skills. Two 20-minute writing tasks are defined by a short prompt that identifies a specific hypothetical situation and audience. The hypothetical situation involves an issue on which the examinee must take a stand. An examinee is instructed to take a position on the issue and to explain to the audience why the position taken is the better (or best) alternative. Two human scorers evaluate each CAAP Essay Writing response; these scores are averaged into a single score for each response.

CAAP Mathematics. The CAAP Mathematics Test is a 35-item, 40-minute test designed to measure students' proficiency in mathematical reasoning. The test assesses students' proficiency in solving mathematical problems encountered in many postsecondary curricula. It emphasizes quantitative reasoning rather than the memorization of formulas. The content areas tested include pre-algebra; elementary, intermediate, and advanced algebra; coordinate geometry; and trigonometry.

CLA

The Collegiate Learning Assessment (CLA), developed and administered by the Council for Aid to Education (CAE), contains three constructed-response tasks: Performance Task, Make-an-Argument, and Critique-an-Argument. For this study, students took either a Performance Task or an Analytic Writing Task (Make-an-Argument plus Critique-an-Argument). The CLA tasks require that students integrate critical thinking, analytic reasoning, problem solving, and written communication skills; therefore, CLA scores reflect multiple constructs. However, for the purposes of this study, the CLA tasks were classified by the constructs that are most essential for successful performance. The Performance Task and Critique-an-Argument task

are classified as critical thinking, although writing skills are also required for both of these measures. The Make-an-Argument task is classified as writing.

CLA Performance Task (CLA PT). The Performance Task is a 90-minute task that requires students to use an integrated set of critical thinking, analytic reasoning, problem-solving, and written communication skills to answer several open-ended questions about a hypothetical yet realistic situation. Each Performance Task has its own document library that includes a range of information sources, such as letters, memos, summaries of research reports, newspaper articles, maps, photographs, diagrams, tables, charts, and interview notes or transcripts. Students are instructed to use these materials in preparing their answers to the Performance Task's questions. Performance Tasks often require students to marshal evidence from different sources; distinguish rational from emotional arguments or fact from opinion; interpret data in tables and figures; deal with inadequate, ambiguous, and/or conflicting information; spot deception and holes in the arguments made by others; recognize information that is and is not relevant to the task at hand; identify additional information that would help to resolve issues; and weigh, organize, and synthesize information from several sources. The CLA utilizes teams of trained and calibrated human scorers to evaluate responses to the Performance Task prompts.

CLA Make-an-Argument (CLA MA). A Make-an-Argument prompt typically presents an opinion on some issue and asks students to write, in 45 minutes, a persuasive, analytic essay to support a position on the issue. Key elements include: establishing a thesis or a position on an issue; maintaining the thesis throughout the essay; supporting the thesis with relevant and persuasive examples (e.g., from personal experience, history, art, literature, pop culture, or current events); anticipating and countering opposing arguments to the position, fully developing ideas, examples, and arguments; crafting an overall response that generates interest, provokes thought, and persuades the reader; organizing the structure of the essay (e.g., paragraphing, ordering of ideas and sentences within paragraphs); employing transitions and varied sentence structure to maintain the flow of the argument; and utilizing sophisticated grammar and vocabulary. The CLA utilizes computer automated scoring to evaluate responses to the Make-an-Argument prompts.

CLA Critique-an-Argument (CLA CA). A "Critique-an-Argument" prompt asks students, in 30 minutes, to critique an argument by discussing how well reasoned they find it to be (rather than simply agreeing or disagreeing with the position presented). Key elements include: identifying a variety of logical flaws or fallacies in a specific argument; explaining how or why the logical flaws affect the conclusions in that argument; and presenting a critique in a written response that is a grammatically correct, organized, well-developed, logically sound, and neutral in tone. The CLA utilizes computer automated scoring to evaluate responses to the Critique-an-Argument prompts.

MAPP

ETS's Measure of Academic Proficiency and Progress (MAPP) is an integrated test that measures college-level proficiency in critical thinking, reading, writing, and mathematics. MAPP offers both a Standard form and an Abbreviated form. In the MAPP Standard form, which was administered for the TVS, each of the four skill areas is measured by 27 multiple-choice items, totaling 108 items. The Standard Form takes about two hours to complete.[2]

MAPP Reading and Critical Thinking. Reading and critical thinking are closely related in the MAPP framework, and for the purpose of proficiency classifications they are combined into a single skill area labeled "reading/critical thinking" (ETS, 2007). For the purpose of this study, however, reading and critical thinking are treated as distinct tests and constructs.

The reading items assess students' ability to recognize factual material explicitly presented in a reading passage, synthesize material from different sections of a passage, identify accurate summaries of the passage or of significant sections of the passage, and discern the main idea or focus of a passage or a significant portion of the passage. Critical thinking items test reading comprehension and critical thinking at the same time. They can be considered as assessing the highest level of reading by asking students to recognize assumptions underlying an argument, evaluate competing hypotheses and explanations, recognize flaws in an argument, and evaluate data for consistency with known facts, hypotheses or methods. These skills are considered critical thinking skills as they involve the use of reasoning and reading skills to engage in analysis and evaluation of written passages.

The reading and critical thinking items are based on information presented in a brief reading selection, picture, graph, etc. Typically, a reading selection serves as the stimulus for one or more reading items and one or more critical thinking items. If the set includes only critical thinking questions, the stimulus may be something other than a reading passage, e.g., a picture or a graph. Each critical thinking and reading item is associated with a humanities, social sciences, or natural sciences academic context.

MAPP Writing. The writing items test grammar and usage, largely in the context of asking students to recognize grammar and usage errors in written material presented to them. The questions measure students' ability to recognize agreement among basic grammatical elements, incorporate new material into a passage, recast existing sentences into new syntactic combinations, discriminate between appropriate and inappropriate use of parallelism, and recognize the most effective revision of a sentence.

MAPP Mathematics. The mathematics questions emphasize arithmetic, especially percents and ratios, algebra, especially translation, and interpreting graphs, tables, etc. There are also some logical reasoning questions. MAPP Mathematics questions aim to assess student ability to interpret and draw inferences from

[2] MAPP can be taken in either paper-and-pencil or online format. Readers are referred to the ETS MAPP Users' Guide (ETS, 2007) for detailed information about MAPP.

mathematical models such as formulas, graphs, tables, and schematics; represent mathematical information symbolically, visually, numerically and verbally; and use arithmetical, algebraic, geometric, and statistical methods to solve problems.

Summary

Table 1 links the constructs involved in this study to the tests described above and their response formats.

Table 1.
Summary of Constructs and Corresponding Tests

Construct(s)	Tests
Critical Thinking	MAPP Critical Thinking, CAAP Critical Thinking, CLA PT*, CLA CA*
Writing	MAPP Writing, CAAP Writing Skills, CAAP WritingEssay*, CLA MA*
Mathematics	MAPP Mathematics, CAAP Mathematics
Reading	MAPP Reading, CAAP Reading
Science	CAAP Science

*Indicates constructed-response test format.

ADMINISTRATIVE METHODS

Student Recruitment

Each of the 13 colleges and universities that participated in the Test Validity Study was responsible for recruiting a sample of 46 freshmen and 46 seniors. Each institution was advised to pre-screen its population of freshmen and seniors for the necessary requirements: at least 18 years of age and SAT/ACT scores on file with the registrar. The freshmen sampled were also required to be full-time, first-time college students. The seniors sampled were required to be students with senior class standing (on track to graduate in spring 2009) and who had initially enrolled as freshmen (non-transfers). Institutions were strongly encouraged to construct a random and representative sample, but it was not required. Most institutions constructed their samples on a first-come, first-served basis.

Each institution was provided with a student information sheet (describing the TVS), a student flyer template, and a student recruitment e-mail template to aid in drawing its sample. Institutions were instructed not to reveal to participating students any details about differences among the tests administered in this study (e.g., the test format). Participation in the study was incentivized at all institutions with a $150 Amazon.com gift certificate in exchange for completion of three separate testing sessions lasting a total of approximately six hours.

TVS Administrative Procedures

The administration of the Test Validity Study took place on 13 campuses between August 2008 and November 2008. TVS administration proctors at each institution were required to complete training via web conference. The 30-minute proctor training web conference covered several topics, including participant recruitment, counterbalancing the order of the tests, participant tracking, supplementary data collection, test administration procedures, and test security.

Participants registered for three TVS sessions (one measure for each session) which took place on three separate days. Participants were permitted to take only one measure per day, but institutions were permitted to offer more than one test session per day to different students. Institutions varied the time and days on which each of the measures were offered such that some students at each institution took CLA, then CAAP, then MAPP; others took CAAP, then MAPP, then CLA, and so forth. Across all institutions, the

order in which students took the three measures was adequately varied, though fewer students took CAAP first and more students took CAAP last than might be expected. [3]

At the beginning of their first TVS session, all participants completed an online demographic survey and a TVS Student Information Card to facilitate proctor tracking of student attendance. All measures were administered under standardized testing conditions, with students beginning at the same time and monitored by proctors. Proctors were instructed to keep the exam room as free from distraction as possible.

At CAAP test sessions, participants took one of two CAAP test packages, each 120 minutes long. The first package consisted of Writing Skills, Mathematics, and Reading modules. The second consisted of Critical Thinking, Science, and Writing Essay modules. Proctors alternated distribution of the two CAAP test packages among participants. CAAP was administered via pencil and paper.

At CLA test sessions, participants took either an Analytic Writing Task (Make-an-Argument plus Critique-an-Argument; 75 minutes) or a Performance Task (90 minutes). The online assessment system randomly distributed tasks among students. All CLA tasks were administered online.

At MAPP test sessions, participants completed the MAPP standard form, which is a two-hour test divided into two 60-minute sections. Each section contained items on critical thinking, reading, writing, and mathematics. MAPP was administered via pencil and paper.

Data Collection

Students completed a five-minute online pre-exam questionnaire for gathering data, including date of birth, gender, race/ethnicity, major, class standing, and transfer status. After the completion of testing at each school, the students' cumulative undergraduate GPAs and admissions test scores (SAT and/or ACT) were obtained from their registrar's office. Test scores on the CLA, MAPP, and CAAP were gathered from CAE, ETS, and ACT, respectively. All student data were de-identified prior to analysis. Student information remained confidential and was not disclosed to anyone outside of the study. Students' results also remained independent from their academic record, and were kept confidential.

Participating Schools and Students

TVS schools were selected to represent a range of institutional characteristics while representing APLU and AASCU member institutions. The 13 TVS schools included two private institutions and 11 public

[3] *Number of students who took each of the measures first and last*

Order	CLA	CAAP	MAPP
First	399	276	400
Last	324	410	341

institutions from different geographic regions, a Historically Black College or University, and two Hispanic-Serving Institutions. Admissions policies ranged from very selective to nearly open admissions. As defined by basic 2005 Carnegie Classification, TVS schools included one baccalaureate college, four master's institutions, and eight research universities:

- Alabama A & M University
- Arizona State University at the Tempe Campus
- Boise State University
- California State University, Northridge
- Florida State University
- Massachusetts Institute of Technology
- Trinity College
- University of Colorado at Denver
- University of Michigan-Ann Arbor
- University of Minnesota-Twin Cities
- University of Texas at El Paso
- University of Vermont
- University of Wisconsin-Stout

Percentages of minority students and women in the study samples aligned with the institutions' overall student bodies. In terms of entering academic abilities, mean SAT or converted ACT scores were slightly higher for the study participants versus their classmates, which is not surprising assuming that students with higher entrance examination scores tend to be more motivated to participate in non-mandatory academic exercises requiring several hours of commitment over multiple days. Please refer to *Appendix B: School and TVS Sample Characteristics* for details regarding the representativeness of the TVS sample of schools and participants.

Participating schools finished testing in late 2008 and collectively met 87 percent of the project goal of 1,200 students. Specifically, 1,051 took all three tests, 23 took only two tests and 51 took only one test. The distribution of students taking all three tests across freshmen (51%) and seniors (49%) was nearly perfect. Please refer to *Appendix C: Summary of File Linkage Procedures* for detail on the construction of the final sample.

Across schools, eight exceeded their sample size goals, three fell just short of their goals, one met 75 percent of its goal, and one met only 25 percent of its goal. This last school faced significant internal challenges in receiving Institutional Review Board approval and operated on a severely truncated timeline.

The online testing window was extended from October 31 through November 24 to give this school and others the opportunity to test more students.

Students were asked to take three tests: CAAP (either the first or second combination of CAAP test modules), CLA (either the Analytic Writing Task or the Performance Task), and MAPP. Owing to this configuration of test packages, which emphasized simplicity from an administrative and test-taker perspective, we exceeded some sample size goals substantially and fell short slightly on others. Please refer to *Appendix D: TVS Testing Counts by Test and Class* for additional detail.

STATISTICAL METHODS

Research Question 1: Correlations among Test Scores

The first research question of this study concerns whether scores from tests that purport to measure the same construct (critical thinking, writing, etc.) and employ the same response format (multiple-choice or constructed-response) are correlated higher with each other than with tests that measure different constructs and/or employ a different response format. Correlation coefficients, which fall in the range of -1.00 to +1.00, reflect the strength of the linear relationship between scores on different tests. A high positive correlation between two tests indicates that students (or schools) who obtain a high score on one test are likely to obtain a high score on the other test. To address the first research question, student-level and school-level correlations among the TVS measures were computed according to the procedures outlined in subsequent sections.

This portion of the TVS sought evidence of convergent and discriminant validity. Evidence of convergent validity is obtained when a test has high correlations with other measures of the same (or a similar) construct. Evidence of discriminant validity is obtained when a test has lower correlations with measures of different constructs than it has with tests assessing the same construct. Such evidence helps confirm that test scores reflect a particular construct (and not other constructs). Note that two tests measuring the same construct should be highly correlated, but a high correlation between two tests does not mean that they measure the same construct. It means only that students with the skills required to perform well on one test tend to have the skills required to perform well on the other test.

Student-level Correlations. Student-level correlations were computed using data from all students who had at least two valid scores on any of the TVS measures, ignoring the grouping of students within classes (freshman and senior) and within schools. Initially, the correlations were computed separately for freshmen and seniors, but differences between the freshman and senior correlations were generally very small (i.e., the correlations were not sensitive to class; see *Appendix E: Differences in Freshman to Senior Correlations*).[4] Thus, it was reasonable to combine the freshman and senior data, which afforded greater precision when estimating the correlations because of increased sample sizes. It was not possible to obtain a correlation for every combination of TVS measures because, by design, not all combinations were administered (e.g., no students completed CLA Performance Task and CLA Make-an-Argument as part of this study).

[4] In addition, we found that the correlation between two tests when based on the pooled freshmen and senior data was usually only 0.01 higher than the average of the corresponding freshmen and senior correlations. Thus, pooling the freshmen and senior data to compute the correlations in Table 2a had virtually no effect on the outcome.

School-level Correlations. Schools are frequently the unit of analysis for institutional assessment (a common use of TVS measures), so school-level correlations were computed to estimate the strength of the linear relationships among school mean scores. These correlations were computed separately using freshman class means and senior class means, and the two were averaged. This method was employed because the freshmen means are not independent of the senior means. It was expected that the pattern of school-level correlations would mirror the pattern of student-level correlations, but school-level correlations are generally higher than student-level correlations because school means are more reliable than individual student scores. If a school-level correlation is very high (approaching 1.00), one could conclude that it does not matter which test is administered for the purpose of rank ordering schools. However, it should be reiterated that a high correlation does not mean that tests measure the same construct.

Research Question 2: Effect Sizes

As tests of college learning, the TVS measures must be sensitive to the growth in skills that occurs over the course of college. Some measures may be more sensitive than others, but comparisons between measures are complicated by the fact that they report scores on different scales. For example, an increase of 25 CLA scale points is fairly small, but an increase of five MAPP scale points is quite large. To address this, standardized difference measures known as effect sizes were computed. In the context of this study, effect sizes reflect the average difference between freshmen and seniors in standard deviation units, adjusted for the difference in freshman and senior mean SAT/ACT scores. The procedures described below for computing effects sizes were repeated for each TVS measure.

The 13 school effect sizes were combined to obtain a precision-weighted composite effect size (d+) (see F1 in *Appendix F*). When combining effect sizes across schools, this procedure gives more weight to schools with larger sample sizes because greater effect size precision is obtained with larger samples. The standard error of the composite effect size was used to generate a 95% confidence interval (see F2 in *Appendix F*). The width of this interval reflects the variation one would expect to observe if the TVS were repeated many times with different samples of students. Such intervals facilitate interpretations by drawing attention to uncertainty in the estimated effect sizes and by indicating credible differences between effect sizes.

Unfortunately, positive freshman-senior effect sizes are possibly confounded with differences in prior academic achievement. Prior experience reveals that participating seniors tend to be a more selective group than participating freshmen (as indicated by mean SAT or ACT scores), so a difference between freshmen and seniors (i.e., a positive effect size) may reflect a difference in prior academic achievement rather than learning that took place during college. For example, consider that seniors scored an average of 0.48 standard deviations higher than freshmen on the CAAP Critical Thinking test (d+ = 0.48). However, some

portion of this effect size can be accounted for by the 29-point average difference on the SAT between the freshmen and seniors who took the CAAP Critical Thinking test (an effect size of 0.16 using the procedures described above). By subtracting the SAT effect size from the TVS measure effect size, one obtains an interpretable, standardized measure of the freshman-senior difference: the adjusted effect size (d+, adj), 0.32 in this sample. The standard errors of the two effect sizes were combined to estimate the standard error of the adjusted effect sizes, and these were used to generate 95% confidence intervals for the adjusted effect sizes (see F3 in *Appendix F*).

Research Question 3: Reliability

In order to make a valid test score interpretation, one must be confident that the score would have been similar had the test been administered under other acceptable testing conditions. In the field of measurement, this defines the notion of test score reliability. Reliability is indexed by coefficients that range from 0.0 to 1.0, with 1.0 reflecting perfectly reliable measurement (i.e., no random error in the test scores). In many testing programs, student score reliability is of primary concern because consequential decisions are based on the scores (e.g., selection, placement, or certification). The most commonly used student-level reliability coefficient is coefficient alpha, which reflects positive correlations among test items. Student-level reliability coefficients were not computed for this study.

School-level score reliability is of concern when tests are used for institutional assessment (as are most TVS measures). As sample sizes increase, school-level scores (i.e., mean scores) become much more reliable than student-level scores because sampling error decreases. School-level reliability was estimated for the TVS measures using a modified version of the split-sample approach described by Klein, Benjamin, Shavelson, and Bolus (2007). This procedure involves randomly splitting the freshmen in each school into Sample A and Sample B, computing the Sample A and Sample B mean scores at each school, and correlating the Sample A and Sample B means across the 13 schools in the sample. The same procedure was carried out for senior data. The school-level reliability estimates provided by this procedure have two major limitations: they are overly conservative due to the use of half-size samples, and they may have come out differently had a different random splitting been used. A Spearman-Brown correction was used to adjust for the use of half-size samples (see F4 in *Appendix F*), and the mean of 1,000 random splits was computed in order to obtain a stable estimate of the expected value of school-level reliability.

RESULTS

Research Question 1: Correlations among Test Scores

The first TVS research question concerned the correlations among the TVS measures. For reference, Tables 2a and 2b provide complete matrices of student-level and school-level correlations, respectively. Data from these tables were reorganized to facilitate a search for evidence of convergent and discriminant validity. Provided in *Appendix G: Lowest and Highest Available Correlations*, Table G1 documents the three lowest and three highest student-level correlations for each TVS measure and Table G2 displays analogous school-level correlations. If a test measures the construct it purports to measure, one should see that it bears high correlations with tests of similar constructs and lower correlations with tests of different constructs. For example, as one might expect, MAPP Mathematics correlates most highly with CAAP Mathematics and least with CAAP Writing Essay. This indicates that MAPP Mathematics and CAAP Mathematics may measure similar constructs and that MAPP Mathematics and CAAP Writing Essay may measure different constructs. Note that the lowest correlation for nearly all the measures is with CAAP Writing Essay because of this test's apparently low reliability. Thus, it is often more informative to look at the second and third lowest correlations.

Table 2a.

Student-level correlation matrix with standard correlations shown above the diagonal

Construct(s)	Test	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
Critical Thinking	1. MAPP		0.75	0.53	0.52	0.76	0.45	0.68	0.34	0.63	0.46	0.86	0.76	0.74
	2. CAAP			0.58	0.47	0.66	0.39	-	0.32	0.57	-	0.71	-	0.74
	3. CLA PT				-	0.50	-	0.49	0.32	0.46	0.40	0.55	0.52	0.52
	4. CLA CA					0.48	0.47	0.49	0.40	0.46	0.44	0.49	0.50	0.50
Writing	5. MAPP						0.44	0.72	0.33	0.60	0.51	0.73	0.70	0.63
	6. CLA MA							0.44	0.37	0.40	0.39	0.43	0.46	0.39
	7. CAAP								-	0.58	0.48	0.70	0.71	-
	8. CAAP Ess.									0.29	-	0.31	-	0.28
Mathematics	9. MAPP										0.76	0.60	0.55	0.71
	10. CAAP											0.46	0.44	-
Reading	11. MAPP												0.76	0.70
	12. CAAP													-
Science	13. CAAP													

Table 2b.

School-level correlation matrix with standard correlations shown above the diagonal and reliabilities shown on the diagonal

Construct(s)	Test	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
Critical Thinking	1. MAPP	0.93	0.93	0.83	0.93	0.96	0.85	0.89	0.62	0.95	0.93	0.96	0.82	0.93
	2. CAAP		0.87	0.79	0.87	0.94	0.79	0.91	0.75	0.90	0.86	0.93	0.76	0.95
	3. CLA PT			0.75	0.73	0.84	0.67	0.77	0.58	0.91	0.91	0.90	0.76	0.86
	4. CLA CA				0.85	0.92	0.90	0.90	0.61	0.82	0.77	0.91	0.91	0.79
Writing	5. MAPP					0.91	0.86	0.97	0.70	0.92	0.90	0.96	0.87	0.90
	6. CLA MA						0.84	0.83	0.67	0.74	0.72	0.82	0.86	0.69
	7. CAAP							0.88	0.74	0.83	0.78	0.93	0.89	0.81
	8. CAAP Ess.								0.75	0.57	0.56	0.62	0.71	0.61
Mathematics	9. MAPP									0.94	0.98	0.94	0.71	0.98
	10. CAAP										0.92	0.91	0.70	0.96
Reading	11. MAPP											0.91	0.86	0.91
	12. CAAP												0.88	0.65
Science	13. CAAP													0.92

On the whole, patterns of student-level correlations revealed that the TVS measures correlated most highly with measures of similar constructs (e.g., critical thinking correlating with critical thinking, writing with writing, reading with reading, and math with math). Note that, according to the MAPP assessment framework, Critical Thinking and Reading should be treated as though they measure the same construct because comprehending text at a high level requires thinking critically about it. MAPP Reading and Critical Thinking items are often based on the same passage, and this helps explain why their correlation is the highest of all. The lowest correlations were with open-ended measures of writing skills (CAAP Writing Essay, CLA MA, and CLA CA); this was likely related to the lower reliability of brief constructed-response tests (tests including multiple essays would be more reliable). CAAP Mathematics and MAPP Mathematics were also listed frequently among the lowest correlations with other tests, which is not surprising given the obvious difference between mathematics skills and most other constructs.

The problem of confounding low correlations with low reliability is reduced by studying the school-level correlations. This is evidenced by the fact that CLA MA and CLA CA, which are believed to be relatively unreliable at the student level, appear in several lists of the three highest school-level correlations (See Appendix G). Nevertheless, the lists do not change dramatically from the student level to the school

level. Further, the evidence of convergent and discriminant validity from the school-level correlations is limited by the small range between the lowest and the highest correlations. The school-level correlations were generally quite high (around 0.90), which suggests that schools with students who have the skills to perform well on one test tend to have students with the skills to perform well on other tests.

To evaluate the simultaneous effects of construct and response format on the correlations, average correlations with other measures were computed and arranged in Tables 3a (student-level) and 3b (school-level). The strength of these effects may be detected by studying differences between the columns. As expected, the highest correlations appear in the "same construct, same format" column, and the lowest correlations tend to appear in the "different construct, different format" column. Comparing the first and third data columns provides an indicator of the effect of construct (holding response format constant). The differences between these columns are usually not large, but they would be larger if MAPP Critical Thinking and Reading were treated as measures of similar constructs (as for the reasons discussed above). The effect of construct might also be observed by comparing the second and fourth data columns, but the averages in those columns are often based on few correlations since there are fewer constructed-response tests and student-level reliabilities of those tests are believed to be relatively low. For the same reasons, discerning the effect of format is difficult, but comparisons between the first and second data columns and the third and fourth columns indicate that "same format" correlations tend to be higher than "different format" correlations. The same data patterns were observed in Table 3b for average school-level correlations. Note that average correlations increase slightly by excluding CAAP Writing Essay from the analysis (values shown in parentheses).

Table 3a.
Average student-level correlations organized by similar/different construct and similar/different response format. Correlations shown in parentheses exclude CAAP-Ess. The -- symbol indicates that correlations could not be computed.

Construct(s)	Test	Same construct Same format	Same construct Diff. format	Diff. construct Same format	Diff. construct Diff. format
Critical Thinking	1. MAPP	0.75	0.53	0.71	0.40 (0.45)
	2. CAAP	0.75	0.53	0.68	0.36 (0.39)
	3. CLA PT	--	0.56	0.32 (--)	0.49
	4. CLA CA	--	0.50	0.44 (0.47)	0.48
Writing	5. MAPP	0.72	0.39 (0.44)	0.66	0.49
	6. CLA MA	0.37 (--)	0.44	0.47	0.42
	7. CAAP	0.72	0.44	0.64	0.49
	8. CAAP Ess.	0.37	0.33	0.36	0.31
Mathematics	9. MAPP	0.76	--	0.61	0.40 (0.44)
	10. CAAP	0.76	--	0.47	0.41
Reading	11. MAPP	0.76	--	0.70	0.45 (0.49)
	12. CAAP	0.76	--	0.65	0.49
Science	13. CAAP	--	--	0.71	0.43 (0.47)

Table 3b.
Average school-level correlations organized by similar/different construct and similar/different response format. Correlations shown in parentheses exclude CAAP-Ess. The -- symbol indicates that correlations could not be computed.

Construct(s)	Test	Same construct Same format	Same construct Diff. format	Diff. construct Same format	Diff. construct Diff. format
Critical Thinking	1. MAPP	0.94	0.93	0.95	0.82 (0.88)
	2. CAAP	0.94	0.84	0.89	0.72 (0.77)
	3. CLA PT	0.83	0.86	0.80 (0.84)	0.89
	4. CLA CA	0.83	0.92	0.86 (0.91)	0.91
Writing	5. MAPP	0.96	0.82 (0.88)	0.95	0.93
	6. CLA MA	0.85 (--)	0.88	0.88	0.84
	7. CAAP	0.96	0.84 (0.88)	0.90	0.91
	8. CAAP Ess.	0.85	0.76	0.78	0.70
Mathematics	9. MAPP	0.95	--	0.93	0.81 (0.86)
	10. CAAP	0.95	--	0.84	0.75 (0.80)
Reading	11. MAPP	0.91	--	0.95	0.90 (0.93)
	12. CAAP	0.91	--	0.83	0.91 (0.92)
Science	13. CAAP	--	--	0.92	0.79 (0.84)

Research Question 2: Effect Sizes

The results presented here include observed and adjusted effect sizes reflecting the average difference between participating freshmen and seniors on the TVS measures. The observed (unadjusted) effect sizes and their corresponding 95% confidence intervals provided in Table 4a (and displayed in Figure 1a) indicate that there were significant differences between the freshmen and seniors on all measures except CAAP Mathematics. Recall, however, that some component of the positive effect sizes reflects differences in entering ability rather than learning that took place during college. Across the TVS measures (excluding CAAP Mathematics), the average effect size was 0.42, and the average difference in ability between freshmen and seniors (as measured by the SAT or ACT) reflected an effect size of 0.10. This suggests that 24% (.10/.42) of the observed freshman-senior difference can be accounted for by entering ability differences.

Adjusted effect sizes, which control for differences in entering ability, are provided in Table 4b and displayed in Figure 1b. The adjustment tends to make the effect sizes smaller and the 95% confidence intervals larger. Although three adjusted effect sizes were not significantly different from zero (CLA Performance Task, CAAP Writing Essay, and CAAP Mathematics), all adjusted effect size estimates were positive except for CAAP Mathematics, which indicates that the TVS measures are sensitive to the increase in skills that occurs over the course of college. The largest adjusted effect sizes were 0.46 for MAPP Critical Thinking, 0.46 for CAAP Reading, 0.45 for MAPP Reading, and 0.40 for CLA Critique-an-Argument. Figure 1b shows that the confidence intervals for all positive adjusted effect sizes overlap somewhat, and this suggests that many differences in adjusted effect sizes were not statistically significant. This was especially true of the writing tests, which had adjusted effect sizes ranging from 0.22 to 0.32. The MAPP and CAAP Reading tests also had very similar adjusted effect sizes (0.45 and 0.46, respectively). There was greater variation among the tests that measure critical thinking skills.

Table 4a.

Precision-weighted average observed effect sizes

Measure	d_+	$se(d_+)$	95% Conf. Interval Lower	Upper
MAPP Critical Thinking	0.57	0.064	0.44	0.69
CAAP Critical Thinking	0.48	0.091	0.30	0.65
CLA Performance Task	0.47	0.090	0.30	0.65
CLA Critique-an-Argument	0.39	0.090	0.22	0.57
MAPP Writing	0.34	0.063	0.22	0.46
CLA Make-an-Argument	0.28	0.089	0.10	0.45
CAAP Writing Skills	0.36	0.090	0.18	0.54
CAAP Writing Essay	0.37	0.092	0.19	0.55
MAPP Mathematics	0.32	0.063	0.19	0.44
CAAP Mathematics	-0.12	0.089	-0.29	0.06
MAPP Reading	0.55	0.064	0.42	0.67
CAAP Reading	0.48	0.091	0.31	0.66
CAAP Science	0.49	0.091	0.31	0.67

Figure 1a.

Precision-weighted average observed effect sizes

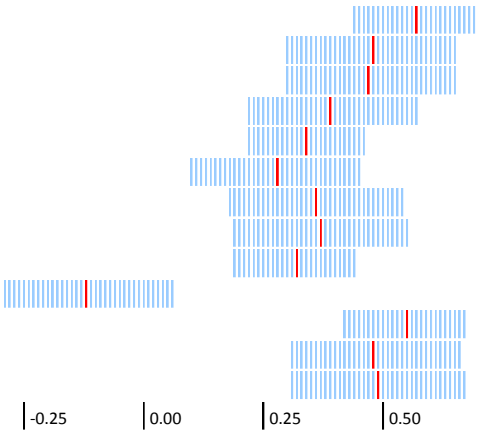


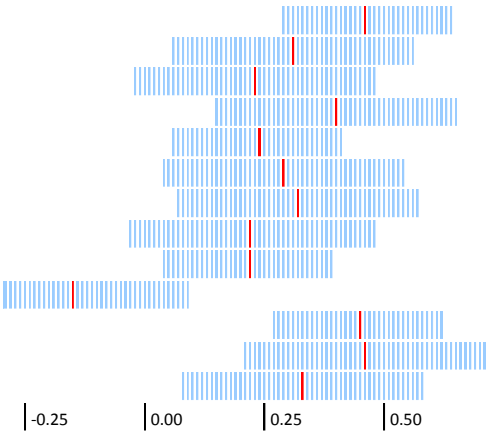
Table 4b.

Precision-weighted average adjusted effect sizes

Measure	$d_{+,adj}$	$se(d_{+,adj})$	95% Conf. Interval Lower	Upper
MAPP Critical Thinking	0.46	0.089	0.29	0.64
CAAP Critical Thinking	0.31	0.128	0.06	0.56
CLA Performance Task	0.23	0.127	-0.02	0.48
CLA Critique-an-Argument	0.40	0.126	0.15	0.65
MAPP Writing	0.24	0.089	0.06	0.41
CLA Make-an-Argument	0.29	0.126	0.04	0.54
CAAP Writing Skills	0.32	0.127	0.07	0.57
CAAP Writing Essay	0.22	0.130	-0.03	0.48
MAPP Mathematics	0.22	0.089	0.04	0.39
CAAP Mathematics	-0.15	0.127	-0.40	0.09
MAPP Reading	0.45	0.089	0.27	0.62
CAAP Reading	0.46	0.129	0.21	0.71
CAAP Science	0.33	0.128	0.08	0.58

Figure 1b.

Precision-weighted average adjusted effect sizes



Research Question 3: Reliability

Recall that a test score must be reliable in order to serve as a valid indicator of a student's or school's performance level. Table 5 provides a summary of school-level reliability coefficients for the TVS measures. The school-level reliability coefficients indicate that scores from these tests are adequately reliable by most standards. A few coefficients are smaller than would typically be observed, but these anomalous values may simply reflect instability of estimates in the small sample of colleges. Generally, the school-level reliabilities were high (greater than 0.90), and this bodes fairly well for the use of relatively small samples for institutional assessment. The within-school sample sizes never exceeded 50 students for MAPP and never exceeded 30 for

CLA or CAAP. It should be noted, however, that the between-school variance was quite large given the small number of schools, which would have a positive impact on school-level reliability.

Table 5.
School-level reliabilities computed as the mean of 1,000 random Spearman-Brown adjusted split-half reliabilities

Measure	Freshman	Senior
MAPP Critical Thinking	0.95	0.91
CAAP Critical Thinking	0.86	0.88
CLA Performance Task	0.85	0.64
CLA Critique-an-Argument	0.86	0.84
MAPP Writing	0.94	0.88
CLA Make-an-Argument	0.87	0.81
CAAP Writing Skills	0.92	0.84
CAAP Writing Essay	0.68	0.82
MAPP Mathematics	0.95	0.93
CAAP Mathematics	0.93	0.90
MAPP Reading	0.94	0.88
CAAP Reading	0.92	0.83
CAAP Science	0.92	0.92

CONCLUSIONS

As discussed previously, a "construct" is a type of knowledge, skill or ability that a test is designed to measure. This study relied on multiple-choice tests to assess three constructs (reading, mathematics, and science). Two other constructs (critical thinking and writing) were measured with both multiple-choice and constructed-response tests.

Some analyses were run using student-level data because test results are often used to make decisions about individuals such as identifying areas where they need remediation or whether they are ready to move on to more challenging courses. School-level analyses were also conducted because test results may be used to inform policy, resource allocation, and programmatic decisions. For example, school-level results may indicate whether the progress students are making at a college is commensurate with the progress made by students at other schools or whether the progress within a school in one domain (e.g., writing) is greater than it is in other areas.

The research described in this report was conducted to answer three main questions about three commonly used tests of student learning for college students: CAAP, CLA, and MAPP. These questions and the answers to them are summarized below.

Research Question 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?

In this study, the pattern of correlations among measures generally supported their construct validity. In other words, results were consistent with the conclusion that tests purporting to measure the same or similar constructs do indeed measure those constructs (and not other constructs). Specifically, an examination of the student-level correlations revealed that two tests of the same construct usually correlated higher with each other than they did with measures of other constructs provided the response format was taken into consideration. For example, when students were the unit of analysis, the average correlation between multiple-choice tests of reading ($r = .76$) [5] was higher than the average correlation between all multiple-choice tests of different constructs ($r = .65$). Similarly, the average correlation between multiple-choice and constructed-response tests of critical thinking ($r = .53$) was higher than it was between multiple-choice and constructed-response tests of different constructs ($r = .45$).

There were two noticeable exceptions to these trends. First, there was an especially low correlation between multiple-choice and constructed-response tests of writing ($r = .40$ on average). This finding may

[5] This correlation, and those that follow, reflect averages of Fisher Z-transformed correlations.

stem from the lower score reliability of constructed-response tests relative to multiple-choice exams, from these measures assessing somewhat different constructs, or from some combination of these and other factors. Second, there was an especially low correlation between two constructed-response writing tests, namely: CAAP Writing Essay and CLA Make-an-Argument (presumably due to low reliability). As discussed below, the correlations between these measures increased dramatically when the school (rather than the student) was the unit of analysis.

The pattern of results at the school level was much fainter because all the correlations were much higher and the differences among them much smaller. This came about as a result of the much higher level of score reliability for all the measures at the school level. For example, the mean correlation between two multiple-choice tests of the same construct (r = .94) at the school level was only very slightly higher than the mean correlation between two multiple-choice tests of different constructs (r = .92). The mean correlation between two constructed-response tests of the same construct (r = .84) at the school level was only slightly higher than the mean correlation between two constructed-response tests of different constructs (r = .83). In addition, the mean correlation between multiple-choice and constructed-response tests of critical thinking (r = .89) was only slightly higher than it was between constructed-response and multiple-choice tests of different constructs (r = .85) or among constructed-response tests of different constructs (r = .83). There also continued to be a lower correlation between multiple-choice and constructed-response tests of writing (r = .83). Thus, while there was less differentiation among the coefficients, the pattern of results at the school level was consistent with the pattern at the student level.

The high correlations among all the tests at the school level may suggest to some that they all measure the same general ability. However, high correlations also can occur if the students who are proficient in one area also tend to be proficient in the other areas that are tested. For instance, it is evident from a side-by-side comparison of the CAAP Mathematics and Reading tests that these tests measure different skills. High correlations also can occur if some of the skills tested, such as reading, are prerequisites for other skills, such as writing. In addition, high correlations say nothing about the students' level of performance or what skills they need to develop to improve their scores.

Finally, what skills are tested and how they are assessed affects instruction. For example, devoting some testing time to open-ended measures in a test battery sends a message that students need to be able to generate ideas and communicate effectively in writing – and this is bound to affect instruction and the types of tests faculty use in their courses.

Research Question 2: Is the difference in average scores between freshmen and seniors related to the construct tested, response format, or the test's publisher?

The first step in answering this question involved creating an index (the "adjusted" effect size) that allowed for comparing score gains between freshmen and seniors in a way that controlled for differences in score distributions (i.e., means and standard deviations) 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 freshman and senior classes.

With one notable exception (discussed below), effect sizes ranged from approximately one quarter to one half of a standard deviation. Furthermore, effect sizes were fairly consistent across tests, test formats (multiple-choice and constructed-response), test publishers (ACT, CAE, and ETS), and constructs. The only conspicuous exception to these trends occurred on the CAAP Mathematics test. The seniors who took this test earned a slightly lower mean score than freshmen. It is not clear why this occurred, although one hypothesis (offered by ACT) is that the test has too low a ceiling; although this ceiling could not by itself explain a negative gain, it could explain a very small gain and other factors could turn such a gain negative. For example, college freshmen who recently prepared for entrance exams could be more proficient in math than college seniors who had not used their math skills recently. This hypothesis is undermined by the reversal only occurring on CAAP Mathematics; it did not occur on MAPP Mathematics.

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

Reliability refers to score consistency. Reliability is typically reported on a scale from 0.00 to 1.00 where higher values indicate more reliable scores.

Practical constraints on the research design precluded our gathering the data that would be needed to compute score reliability at the student level for the four constructed-response tests, but past research suggests that those coefficients are likely to be considerably lower than those obtained with the multiple-choice measures (see, e.g., Klein & Bolus, 2003). This difference in reliability is no doubt a major source of the relatively low student-level correlations between the constructed-response and multiple-choice tests discussed below. However, score reliability was high when the school was used as the unit of analysis, ranging from .75 to .84 for the constructed-response tests, and from .88 to .94 for the multiple-choice tests. Thus, score reliability is not a serious concern when results are reported at the school level and each school recruits student samples with sizes comparable to those used in this study (30-50 students). Of course, larger samples and truly random samples should still be the goal of any institution seriously interested in value-added assessments.

Summary Conclusions

Overall, across test constructs, response formats, and test publishers: correlations are generally high at the school level, adjusted effect sizes are consistent, and school-level reliabilities are high. At the student level, the correlations provided evidence supporting the construct validity of the tests.

The correlations between scores on different tests, especially at the student level, are affected by the reliability of those scores. When the individual student is the unit of analysis, multiple-choice tests are known to yield more reliable scores per hour of testing time than constructed-response measures. This is one explanation for the fact that correlations tended to be lower when one or both tests employed a constructed-response format (it may also be that they measure different constructs). This suggests 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 should be limited to instances where the differences are statistically significant.

In contrast, when the analysis is conducted at the school level, we found that all of the measures have high reliabilities, and the schools' mean scores on different tests all correlate highly regardless of response format or the construct measured by the test. 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. The findings above 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 also could stem from the fact that many students who have the skills needed to achieve in one area also have the skills necessary for other areas.

Finally, given the findings above and particularly the high correlation among the measures, the decision about which measures to use will probably hinge on their acceptance by students, faculty, administrators, and other policy makers. There also may be trade-offs in costs, ease of administration, and the utility of the different tests for other purposes, such as to support other campus activities and services. Indeed, the assessment 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. For this to be accomplished systematically and systemically, adopters of any test covered in this study should also understand the catalytic roles played by campus leadership, willing faculty, and cultures of evidence. Equally important are the benefits inherent in assessment tools that are reliable (correlate well with other tools), have face validity (represent the type of performance you want students to demonstrate), and that couple summative data with formative diagnostics to improve teaching and learning.

REFERENCES

ACT. (2008). *CAAP technical handbook*. Iowa City, IA, US: ACT.

College Board. (2008). *Trends in college pricing (2008)*. New York, NY: College Board.

Educational Testing Service. (2007). *MAPP user's guide*. Princeton, NJ: Educational Testing Service.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review, 31*(5), 415-439.

Klein, S. & Bolus, R. (2003). *Factors affecting score reliability on high stakes essay exams*. Paper presented at the meetings of the American Educational Research Association, Chicago, IL.

U.S. Census Bureau. (2008). *Current population survey*. Washington D.C.: U.S. Census Bureau.

APPENDIX A: SAMPLE ITEMS

CAAP

Sample CAAP Critical Thinking Item:

Directions: There are four passages in this test. Each passage is followed by several questions. After reading a passage, choose the best answer to each question by circling the corresponding answer option. You may refer to the passages as often as necessary.

Senator Favor proposed a bill in the state legislature that would allow pharmacists to prescribe medications for minor illnesses, without authorization from a physician (i.e., a "prescription"). In support of her proposal, Favor argued:

Doctors have had a monopoly on authorizing the use of prescription medicines for too long. This has caused consumers of this state to incur unnecessary expense for their minor ailments. Often, physicians will require patients with minor complaints to go through an expensive office visit before the physician will authorize the purchase of the most effective medicines available to the sick.

Consumers are tired of paying for these unnecessary visits. At a recent political rally in Johnson County, I spoke to a number of my constituents and a majority of them confirmed my belief that this burdensome, expensive, and unnecessary practice is widespread in our state. One man with whom I spoke said that his doctor required him to spend $80 on an office visit for an uncommon skin problem which he discovered could be cured with a $2 tube of prescription cortisone lotion.

Anyone who has had to wait in a crowded doctor's office recently will be all too familiar with the "routine": after an hour in the lobby and a half-hour in the examining room, a physician rushes in, takes a quick look at you, glances at your chart and writes out a prescription. To keep up with the dizzying pace of "health care," physicians rely more and more upon prescriptions, and less and less upon careful examination, inquiry, and bedside manner.

Physicians make too much money for the services they render. If "fast food" health care is all we are offered, we might as well get it at a good price. This bill, if passed into law, would greatly decrease unnecessary medical expenses and provide relief to the sick: people who need all the help they can get in these trying economic times. I urge you to vote for this bill.

After Senator Favor's speech, Senator Counter stood to present an opposing position, stating:

Senator Favor does a great injustice to the physicians of this state in generalizing from her own health care experiences. If physicians' offices are crowded, they are crowded for reasons that are different from those suggested by Senator Favor. With high operating costs, difficulties in collecting medical bills, and exponential increases in the costs of malpractice insurance, physicians are lucky to keep their heads above water. In order to do so, they must make their practices more efficient, relying upon nurses and laboratories to do some of the patient screening.

No one disputes the fact that medical expenses are soaring. But, there are issues at stake which are more important than money—we must consider the quality of health care. Pharmacists are not trained to diagnose illnesses. Incorrect diagnoses by pharmacists could lead to extended illness or even death for an innocent customer. If we permit such diagnoses, we will be personally responsible for those illnesses and deaths.

Furthermore, since pharmacies make most of their money by selling prescription drugs, it would be unwise to allow pharmacists to prescribe. A sick person who has not seen a physician might go into a drugstore for aspirin and come out with narcotics!

Finally, with the skyrocketing cost of insurance, it would not be profitable for pharmacists to open themselves up to malpractice suits for mis-prescribing drugs. It is difficult enough for physicians with established practices to make it; few pharmacists would be willing to take on this financial risk. I recommend that you vote against this bill.

Favor's "unofficial poll" of her constituents at the Johnson County political rally would be more persuasive as evidence for her contentions if the group of people to whom she spoke had:

I. been randomly selected.
II. represented a broad spectrum of the population: young and old, white and non-white, male and female, etc.
III. not included an unusually large number of pharmacists.

(A) I only
(B) II only
(C) III only
(D) I, II, and III

Sample CAAP Science Item (Biology, Data Representation Format):

Directions: There are eight passages in this test. Each passage is followed by several questions. After reading a passage, choose the best answer to each question by circling the corresponding answer option. You may refer to the passages as often as necessary.

A scientist investigated the factors that affect seed mass in the plant species Desnodium poniculatum. Some results of this study are summarized in the two tables below.

Table 1

Daylight hours	Other variable	Average seed mass (in mg) of plants raised at:	
		23ºC	29ºC
14	—	7.10	5.63
14	Leaves removed	7.15	6.11
14	Reduced water	4.81	5.81
8	—	6.12	—

Table 2

A. Number of seeds per fruit	Average seed mass (mg)
1 2 3 4 5	6.62 6.28 5.97 6.00 5.59
B. Position of seed in fruit*	Average seed mass (mg)

1 (closest to stem)	5.98
2	6.06
3	5.96
4	5.82
5 (farthest from stem)	5.27

*Seeds closest to the stem mature first and are released first.

The data suggest that subjecting plants to which of the following conditions would result in the greatest seed masses?

(A) 8 hours of light, adequate water supply, and 23ºC
(B) 8 hours of light, decreased water supply, and 23ºC
(C) 14 hours of light, adequate water supply, and 23ºC
(D) 14 hours of light, decreased water supply, and 29ºC

Sample CAAP Reading Item:

Directions: There are four passages in this test. Each is followed by nine questions. After reading a passage, choose the best answer to each question by circling the corresponding answer option. You may refer to the passages as often as necessary.

Passage 2 Prose Fiction

On Union Boulevard, St. Louis, in the 1950's, there were women in their eighties who lived with the shades drawn, who hid like bats in the caves they claimed for home. Neighbors of my grandmother, they could be faintly heard through a ceiling or wall. A drawer opening. The slow thump of a shoe. Who they were and whom they were mourning (someone had always just died) intrigued me. Me, the child who knew where the cookies waited in Grandma's kitchen closet. Who lined five varieties up on the table and bit from each one in succession, knowing my mother would never let me do this at home. Who sold Girl Scout cookies door-to-door in annual tradition, who sold fifty boxes, who won The Prize. My grandmother told me which doors to knock on. Whispered secretly, "She'll take three boxes—wait and see."

Hand-in-hand we climbed the dark stairs, knocked on the doors. I shivered, held Grandma tighter, remember still the smell which was curiously fragrant, a sweet soup of talcum powder, folded curtains, roses pressed in a book. Was that what years smelled like? The door would miraculously open and a withered face framed there would peer oddly at me as if I had come from another world. Maybe I had. "Come in," it would say, or "Yes?" and I would mumble something about cookies, feeling foolish, feeling like the one who places a can of beans next to an altar marked For the Poor and then has to stare at it—the beans next to the cross—all through the worship. Feeling I should have brought more, as if I shouldn't be selling something to these women, but giving them a gift, some new breath, assurance that there was still a child's world out there, green grass, scabby knees, a playground where you could stretch your legs higher than your head. There were still Easter eggs lodged in the mouths of drainpipes and sleds on frozen hills, that joyous scream of flying toward yourself in the snow. Squirrels storing nuts, kittens being born with eyes closed; there was still everything tiny, unformed, flung wide open into the air!

But how did you carry such an assurance? In those hallways, standing before those thin gray wisps of women, with Grandma slinking back and pushing me forward to go in alone, I didn't know. There was something here which also smelled like life. But it was a life I hadn't learned yet. I had never outlived anything I knew of, except one yellow cat. I never had saved a photograph. For me life was a bounce, an unending burst of pleasures. Vaguely I imagined what a life of recollection could be, as already I was haunted by a sense of my own lost baby years, golden rings I slipped on and off my heart. Would I be one of those women?

Their rooms were shrines of upholstery and lace. Silent radios standing under stacks of magazines. Did they work? Could I turn the knobs? Questions I wouldn't ask here. Windows with shades pulled low, so the light peeping through took on a changed quality, as if it were brighter or dimmer than I remembered. And portraits, photographs, on walls, on tables, faces strangely familiar, as if I was destined to know them. I asked no questions and the women never questioned me. Never asked where the money went, had the price gone up since last year, were there any additional flavors. They bought what they remembered—if it was peanut-butter last year, peanut-butter this year would be fine. They brought the coins from jars, from pocketbooks without handles, counted them carefully before me, while I stared at their thin crops of knotted hair. A Sunday brooch pinned loosely to the shoulder of an everyday dress. What were these women thinking of?

And the door would close softly behind me, transaction complete, the closing click like a drawer sliding back, a world slid quietly out of sight, and I was free to return to my own universe, to Grandma standing with arms folded in the courtyard, staring peacefully up at a bluejay or sprouting leaf. Suddenly I'd see Grandma in her dress of tiny flowers, curly gray permanent, tightly laced shoes, as one of them—but then she'd turn, laugh, "Did she buy?" and again belong to me.

Gray women in rooms with the shades drawn . . . weeks later the cookies would come. I would stack the boxes, make my delivery rounds to the sleeping doors. This time I would be businesslike, I would rap firmly, "Hello Ma'am, here are the cookies you ordered." And the face would peer up, uncertain . . . cookies? . . . as if for a moment we were floating in the space between us. What I did (carefully balancing boxes in both my arms, wondering who would eat the cookies—I was the only child ever seen in that building) or what she did (reaching out with floating hands to touch what she had bought) had little to do with who we were, had been, or ever would be.

Naomi Shihab Nye, "The Cookies." © 1982 by Naomi Shihab Nye.

Which of the following statements represents a justifiable interpretation of the meaning of the story?

(A) The girl's experience selling Girl Scout cookies influenced her choice of careers.
(B) The girl's experiences with elderly women made her aware of the prospect of aging.
(C) Because she spent so much time with her grandmother, the girl preferred the company of older people to that of other children.
(D) The whole experience of selling Girl Scout cookies was a dream or hallucination and had nothing to do with who the girl really was.

Sample CAAP Writing Skills Item:

Directions: In the six passages that follow, certain words and phrases are underlined and numbered. In the right-hand column, you will find alternatives for each underlined part. You are to choose the one that best expresses the idea, makes the statement appropriate for standard written English, or is worded most consistently with the style and tone of the passage as a whole. If you think the original version is best, choose "NO CHANGE."

In the end, everyone gives up jogging. Some find that their strenuous efforts to earn a living **drains (1)** away their energy.

(A) NO CHANGE
(B) drain
(C) has drained
(D) is draining

Sample CAAP Writing Essay Prompt:

Your college administration is considering whether or not there should be a physical education requirement for undergraduates. The administration has asked students for their views on the issue and has announced that its final decision will be based on how such a requirement would affect the overall educational mission of the college. Write a letter to the administration arguing whether or not there should be a physical education requirement for undergraduates at your college.
(Do not concern yourself with letter formatting; simply begin your letter, "Dear Administration.")

Sample CAAP Mathematics Item (Pre-Algebra Application):

Directions: Solve each problem, then choose the correct answer by circling the corresponding answer option. Do not linger over problems that take too much time. Solve as many as you can; then return to the others in the time you have left for this test. You may use a calculator for any of the problems on this test. However, all problems can be solved without using a calculator, and some of the problems may in fact be simpler if done without a calculator.

Mark bought 3 shirts at a clothing store. If he paid a total of $15.00 for 2 shirts and the average (arithmetic mean) cost of the 3 shirts was $8.00, how much did Mark pay for the third shirt?
(A) $7.00
(B) $7.67
(C) $8.50
(D) $9.00
(E) $11.50

MAPP

Sample MAPP Reading and Critical Thinking Items:

Directions: Each stimulus (a passage, poem, graph, or table, for example) is followed by a question or questions based on that stimulus. Read each stimulus carefully. Then choose the best answer to each question following a stimulus.

Certain literary theorists claim to see no difference between literature and criticism. They rest their case on two similarities between the genres: both are impassioned and both use "literary language." The critical essays of John Ruskin (1819--1900) are surely impassioned, and surely full of literary language. However, we do recognize a difference, not in the use of language, but in the internal organization of parts between the literary genres (the novel, drama, poetry), which tend to be organized around a central, defining symbol or set of symbols, and the nonliterary ones (homily, criticism, the philosophical essay), which tend to be linear and discursive in nature. It is by some such structural principle, and not by any remarks about language, that we distinguish the critical essay from literary genres such as poetry.

Reading

The primary purpose of the passage is to
(A) analyze a major trend in recent literary theory
(B) point out the distinguishing features of certain important literary genres
(C) question the claim that there are significant differences between literary and nonliterary genres
(D) identify a means of differentiating between literary and nonliterary genres

Critical Thinking

Which of the following claims, if true, would be most difficult to reconcile with the argument made by the author of the passage?

(A) Few essayists are as skilled in their use of literary language as Ruskin was.
(B) Many prose poets tend to avoid the use of impassioned literary language in their work.
(C) The use of the symbol as a structuring device in poetry is more common in certain literary periods than in others.
(D) The essay form was invented in the late sixteenth century as a way for writers to articulate personal thoughts and feelings.

Sample MAPP Writing Item:

Directions: The following question tests your ability to rewrite a given sentence. You will be told exactly how to revise your new sentence. Keep in mind that your new sentence should have the same meaning as the sentence given to you. In choosing an answer, follow the requirements of standard written English; that is, pay attention to acceptable usage in grammar, diction (choice of words), sentence construction, and punctuation. Choose the best answer; this answer should be clear and exact, without awkwardness, ambiguity, or redundancy.

Being a female jockey, she was often interviewed.

Rewrite, beginning with

She was often interviewed

The next words will be

(A) on account of she was
(B) by her being
(C) because she was
(D) being as she was

Sample MAPP Math Item:

Directions: Solve each problem, using any available space on the page for scratchwork. Then decide which is the best of the choices given and select that answer.

A train traveled at a constant rate of f feet per second. How many feet did it travel in x minutes?

(A) $\frac{60f}{x}$

(B) $\frac{fx}{60}$

(C) $\frac{x}{60f}$

(D) $60fx$

CLA

Sample CLA Performance Task:

Directions: Please read the instructions in Document 1 located in the Document Library (see right side of screen). Your answers to the questions that follow should describe all details necessary to support your position. Your answers will be judged not only on the accuracy of the information you provide, but also on how clearly the ideas are presented, how effectively the ideas are organized, and how thoroughly the information is covered. While your personal values and experiences are important, please answer all questions solely on the basis of the information above and in the Document Library.

You are the assistant to Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of DynaTech's sales force, recommended that DynaTech buy a small private plane (a SwiftAir 235) that she and other members of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation:

1: Newspaper articles about the accident

2: Federal Accident Report on in-flight breakups in single engine planes

3: Pat's e-mail to you & Sally's e-mail to Pat

4: Charts on SwiftAir's performance characteristics

5: Amateur Pilot article comparing SwiftAir 235 to similar planes

6: Pictures and description of SwiftAir Models 180 and 235

Do the available data tend to support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups? What is the basis for your conclusion? What other factors might have contributed to the accident and should be taken into account? What is your preliminary recommendation about whether or not DynaTech should buy the plane and what is the basis for this recommendation?

Sample CLA Make-an-Argument Prompt:

Directions: You will have 45 minutes to plan and write an argument on the topic on the next screen. You should take a position to support or oppose the statement. Use examples taken from your reading, coursework, or personal experience to support your position. Your essay will be evaluated on how well you do the following:
1. State your position
2. Organize, develop, and express your ideas
3. Support your ideas with relevant reasons and/or examples
4. Control the elements of standard written English

Government funding would be better spent on preventing crime than in dealing with criminals after the fact.

Sample CLA Critique-an-Argument Prompt:

Directions: There is something wrong with the argument presented below. It is your job to explain what is wrong with the argument. Discuss any flaws in the argument, any questionable assumptions, any missing information, and any inconsistencies. What we are interested in is your critical thinking skills and how well you write your response. You will have 30 minutes to respond to the argument. You will be judged on how well you do the following:
1. Explain any flaws in the points the author makes
2. Organize, develop, and express your ideas
3. Support your ideas with relevant reasons and/or examples
4. Control the elements of standard written English

The number of marriages that end in divorce keeps growing. A large percentage of them are from June weddings. Because June weddings are so popular, couples end up being engaged for a long time just so that they can get married in the summer months. The number of divorces gets bigger with each passing year, and the latest news is that more than 1 out of 3 marriages will end in divorce. So, if you want a marriage that lasts forever, it is best to do everything you can to prevent getting divorced. Therefore, it is good advice for young couples to have short engagements and choose a month other than June for a wedding.

APPENDIX B: SCHOOL AND TVS SAMPLE CHARACTERISTICS

School	Public	HSI / HBCU	PHD	MA	BA	% Minority	% Women	Mean SAT		% Minority	% Women	Mean SAT	# Freshmen	# Seniors
						IPEDS				TVS Sample				
Alabama A & M University	■	■		■		95	53	865		97	60	1000	44	19
Arizona State University at the Tempe Campus	■		■			18	52	1100		26	48	1187	47	46
Boise State University	■			■		8	54	1005		8	57	1097	46	47
California State University, Northridge	■	■		■		38	59	950		47	54	1005	51	54
Florida State University	■		■			23	57	1160		23	56	1215	46	48
Massachusetts Institute of Technology			■			19	44	1500		31	55	1458	49	47
Trinity College					■	19	50	1310		23	62	1292	43	46
University of Colorado at Denver	■		■			15	57	1025		21	61	1160	18	9
University of Michigan-Ann Arbor	■		■			13	51	1280		12	59	1321	45	47
University of Minnesota-Twin Cities	■		■			7	53	1165		10	37	1256	38	45
University of Texas at El Paso	■	■	■			77	55	*920*		88	54	1022	48	32
University of Vermont	■		■			3	55	1165		7	63	1207	49	46
University of Wisconsin-Stout	■			■		2	50	990		4	54	1131	46	46
	11	3	8	4	1									
Notes:														
IPEDS is Integrated Postsecondary Data System (nces.ed.gov/IPEDS)						26	53	1110	*Mean*	31	55	1181		
For UT El Paso only, Mean SAT is estimated based on CLA takers						28	4	179	*SD*	30	7	134		
Minority excludes White and Asian/Pacific Islander						95	59	1500	*Max*	97	63	1458		
Counts of freshmen and seniors include those who took at least 2 measures						2	44	865	*Min*	4	37	1000		

APPENDIX C: SUMMARY OF FILE LINKAGE PROCEDURES

I. Data File Acquisition

The following six separate data files were received for data processing.

1. "Tracking" File

This data file was created by combining Microsoft Office Excel files received from each of the 13 participating universities. Each incoming file contained a single record for each participating student. On that record, the proctor was to record the 9 digit study id (STUDY ID) in the format cssaaabbb, where c indicated the class (1=Freshman, 4=Senior), ss indicated the school number assigned to the school for the project (01-13), aaa indicated the unique student number, and bbb replicated that student number (e.g., css001001). The tracking file also contained the date of birth of the student, the dates that the student sat for each of the three tests (ACT-CAAP, CAE-CLA, ETS-MAPP), and an indication of which of the two CAAP test packages the student took.

After deletion of extraneous records, the file contained 1,150 records.

2. "Profile" File

Per protocol, all participating students were asked to log into an online testing interface and complete their profile. The system collected the students' names; their dates of birth; the class years they attended their current school (and their current class); their major in school; their race, sex and primary language; and the identification number associated with the CLA test that they took. The system also recorded the STUDY ID entered by the student.

The Profile File contained 1,149 records.

3. "Registrar" File

Utilizing an extract of identification numbers obtained from the online testing interface, the names of participating students were sent to each of the universities registrar offices to obtain information on their admission standardized test scores (ACT and/or SAT), cumulative GPA, class standing, and their transfer

status. The universities each returned an Excel file (along with the STUDY ID) and a single file containing all available records was prepared.

The Registrar File contained 1,148 records.

4. ACT-CAAP File

Per protocol, a subset of students were to have taken one of two ACT-CAAP test batteries and recorded their answers on scanned documents (answer to constructed-response prompts were written in separate booklets). Students were instructed to put their name, date of birth, and STUDY ID on each answer document. Documents were returned to ACT where the multiple-choice tests were scanned and scored, and essay responses were graded. ACT linked the various scored documents such that each student had a single data record prepared in a fix formatted text file.

Per protocol, students were to have taken either CAAP 1 (multiple-choice Writing, Math and Reading) or CAAP 2 (multiple-choice Critical Thinking and Science, and a written Essay). Each data record contained either of these sets of scores, the student's name, date of birth, and STUDY ID. In addition, a vector of individual item scores for each section was provided on the data record.

The incoming ACT-CAAP text file contained 1,113 records.

5. CAE-CLA File

Per protocol, a subset of students were to have taken either the Analytic Writing Task or Performance Task portion of the CAE-CLA online. The Analytic Writing Task portion was composed of two parts, the Make-an-Argument (MA) and Critique-an-Argument (CA). Responses were graded online, and an Excel file was provided to CAE. Each student record in the Excel file included the raw score(s), along with the student's STUDY ID, name, date of birth, task identifier (i.e., which of 8 PT, 4 MA or 4 CA forms were taken), and a vector of individual item scores for the PT. The raw scores on each section were converted to a common scale score by CAE.

The CAE-CLA Excel File contained 1,112 records.

6. ETS-MAPP File

Per protocol, a subset of students were to have taken the ETS-MAPP multiple-choice tests and recorded their answers on test booklets that would be scanned. Students were instructed to put their name and STUDY ID on their answer sheet. Special instructions were to be given by proctors to bubble in digits for month, day and year of birth in a special section of the test booklet, not normally intended for this use.[6]

Scale scores were calculated by ETS for the Mathematics, Reading, Critical Thinking and Writing subtests, which were the primary tests for the study. In addition, scale scores were provided for Humanities, Social Sciences and Natural Sciences topic areas along with an overall total test score.

The ETS-MAPP file contained 1,107 records.

Table C1 summarizes the record count for each of the six data files.

Table C1.
Incoming Project Data Files and Record Counts

Data File	Record Count
Tracking	1,150
Profile	1,149
Registrar	1,148
ACT-CAAP	1,113
CAE-CLA	1,112
ETS-MAPP	1,107

II. File Linkage Method

The end objective of the initial data management step was to create a single record for each study participant. The initial plan was that this record would contain all available student characteristics, test scores and related data. The key linking variable was to be the nine character STUDY ID, with students' date of birth and name as secondary variables. The Profile file was intended to serve as the reference dataset since all students were initially to enter the study through the online system, and here, the STUDY ID would be verified and validated. The Tracking file was to be used as the reference file that would document which combination of test packages the students took. All files would be simultaneously match/merged, and non-linking records investigated and corrected.

[6] The date of birth fields were intended for use in supporting match/merge of files. They were only sporadically completed, and were not used.

For a variety of reasons, the simultaneous linking approach was not feasible.

- The online system was unable to implement an automated validation procedure for the STUDY ID. Thus, many records contained a non-valid STUDY ID (e.g., social security numbers, the student's school identification number, random digits). Additionally, it was learned that a small subset of students did not enter the online system, and did not have a profile record generated.

- The Tracking file was found to have errors caused by mis-recording by proctors.

- Dates of birth were not validated on the files and were either missing and/or inaccurate in many instances.

- Completed files were received over a period of several months, thereby minimizing the efficacy of waiting to do a one-time match merge procedure.

- In part because of the above problems, and the fact that manual, non-validated recording of STUDY ID was required, files had duplicated records and/or different students having the same STUDY ID.

As a result of these issues, it was decided to pursue a sequential, two-step process.

Step 1. As it was received from the field, each individual file was first match/merged with the Profile File using the STUDY ID. Each non-linked, duplicate or missing record problem that occurred, was researched by the respective test publishers. Corrections to the initial files were made by the agencies (e.g., correct STUDY IDs were added to the file and/or changed; dates of births were modified), and the files were resent. The Profile file was then iteratively re-linked with each of the files using both STUDY ID and then, available secondary variables (Date of Birth and Student Names). Corrections were made to the files at each iteration. The objective of each iteration was to get the most valid STUDY ID onto each file.

Step 2. Once it was determined that each file had an accurate and validated STUDY ID, all six files were simultaneously match/merged. This linkage revealed some inconsistencies that were subsequently corrected and a linked file was created.

III. File Linkage Results

The final six-way match resulted in the findings presented in Table C2.

Table C2.
Results of Linkage of All Project Data Files

Description	Record Count
Total Linked Records	1,156
Records containing only 1 publishers' test	54
Records containing only 2 publishers' tests	28
Records containing all 3 publishers' tests	1,074
Records Missing a Profile File	7
Records Missing a Registrar File	8

Per design, all students were to have taken each of the agencies test. Further, students taking only one of the tests were to be dropped from the study since their results could not be correlated with any of the other tests. Thus, the final sample was 1,102 students.

In the final sample of 1,102 students, all records from a test publisher (or the college registrar's office) did not necessarily contain valid scores for each of the measures. This may have been due to a student not taking one of the tests within the package, not completely finishing a task (as determined by the test publishers' scoring criteria), leaving before being presented with the specific task, or in the case of the registrar's office data, the information was not in the files or was not recorded.

Table C3 documents the count of valid scores per measure for each of the key variables in the study.

Table C3.
Study Measure and the Count of Valid, non-missing scores

Measure	Valid Scores
College Registrar	
SAT/ACT Equivalent Score	1064
CAE-CLA	
Performance Task Scaled Score	544
MA Scaled Score	546
CA Scaled Score	546
ACT-CAAP	
Writing Skills Scaled Score	554
Mathematics Scaled Score	552
Reading Scaled Score	541
Critical Thinking Scaled Score	539
Science Scaled Score	538
Composite Essay Score	520
ETS-MAPP	
Total Scaled Score	1092
Critical Thinking Scale Score	1092
Reading Scaled Score	1092
Writing Scaled Score	1092
Mathematics Scaled Score	1092
Humanities Scaled Score	1092
Social Sciences Scaled Score	1092
Natural Sciences Scaled Score	1092

IV. Analysis Extract File

A decision was made to create an extract file that could be used individually by each of the testing agencies for the purpose of replicating analyses. Each of the publishers used SAS software for their analyses and agreed that the extract could remain in this format[7].

Two extract files were prepared. Student names and other identifiers (except the STUDY ID) were removed from the files. The first file contained key demographics and summary scores for each of the students. The

[7] Note that all match/merging, databases and data analysis for the project utilized SAS software.

second file contained similar descriptive information, but scored item response vectors were substituted for the actual summary scores. The files could be linked utilizing the STUDY ID variable.

APPENDIX D: TVS TESTING COUNTS BY TEST AND CLASS

Because each student took the full MAPP package, but only half of the CLA or CAAP tests, we expected to see higher counts in cells involving MAPP and lower counts in cells where the CLA is paired with CAAP. The tables below provide details on these correlation pairings. Testing counts appear on the right and percentages of goal met appear on the left.

TVS Summary Testing Counts

Freshmen	*Min 120*		*Max 285*	
	MA CA	PT	CAAP 1	CAAP 2
MAPP	284	280	285	274
MA CA			135	150
PT			153	120

Seniors	*Min 122*		*Max 269*	
	MA CA	PT	CAAP 1	CAAP 2
MAPP	254	269	264	254
MA CA			127	122
PT			136	132

All Students	*Min 252*		*Max 549*	
	MA CA	PT	CAAP 1	CAAP 2
MAPP	538	549	549	528
MA CA			262	272
PT			289	252

Testing Counts as percentage of goal

Freshmen	*Min 80%*		*Max 190%*	
	MA CA	PT	CAAP 1	CAAP 2
MAPP	189%	187%	190%	183%
MA CA			90%	100%
PT			102%	80%

Seniors	*Min 81%*		*Max 179%*	
	MA CA	PT	CAAP 1	CAAP 2
MAPP	169%	179%	176%	169%
MA CA			85%	81%
PT			91%	88%

All Students	*Min 84%*		*Max 183%*	
	MA CA	PT	CAAP 1	CAAP 2
MAPP	179%	183%	183%	176%
MA CA			87%	91%
PT			96%	84%

Goal per cell: 150 for each class and 300 for all students

APPENDIX E: DIFFERENCES IN FRESHMAN TO SENIOR CORRELATIONS

Difference in Freshman to Senior Correlations (Student Level)

Construct	Test	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Critical Thinking	1. MAPP		-0.05	0.06	0.08	0.02	0.15	-0.02	-0.04	0.09	0.09	-0.02	0.10	0.06
	2. CAAP			0.06	0.11	-0.04	0.20		0.02	0.02		-0.04		-0.04
	3. CLA-PT					0.03		0.17	0.08	0.02	0.15	0.00	0.07	0.04
	4. CLA-CA					0.11	0.09	0.09	0.08	0.22	0.19	0.05	0.18	0.16
2. Writing	5. MAPP-Writing						0.15	0.01	0.03	0.07	-0.03	0.00	0.12	0.09
	6. CLA-MA							0.06	0.15	0.18	0.07	0.15	0.04	0.20
	7. CAAP-Writing									0.08	0.04	0.03	0.06	
	8. CAAP-Essay									-0.06		0.00		0.00
3. Mathematics	9. MAPP										0.00	0.09	0.17	0.07
	10. CAAP											0.08	0.16	
4. Reading	11. MAPP												0.06	0.09
	12. CAAP													
5. Science	13. CAAP													

Difference in Freshman to Senior Correlations (School Level)

Construct	Test	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Critical Thinking	1. MAPP		0.01	-0.02	0.04	0.04	0.17	-0.01	0.48	0.03	0.16	0.00	-0.08	0.01
	2. CAAP			0.01	0.15	0.12	0.27		0.77	0.05		0.06		0.02
	3. CLA-PT					-0.01		0.11	0.16	0.10	0.18	0.03	-0.08	0.01
	4. CLA-CA					0.14	0.26	0.04	0.35	0.06	0.13	0.08	0.03	0.07
2. Writing	5. MAPP-Writing						0.11	0.09	0.43	-0.05	-0.01	0.00	0.03	-0.02
	6. CLA-MA							0.29	0.01	0.12	0.08	0.07	0.10	0.02
	7. CAAP-Writing									-0.04	-0.02	0.09	0.13	
	8. CAAP-Essay									0.40		0.37		0.38
3. Mathematics	9. MAPP										0.07	0.00	-0.11	0.05
	10. CAAP											0.11	-0.04	
4. Reading	11. MAPP												-0.03	0.00
	12. CAAP													
5. Science	13. CAAP													

APPENDIX F: EQUATIONS REFERENCED IN STATISTICAL METHODS AND RESULTS

F1: $d_+ = \sum_i w_i d_i$, where $d_i = \left(1 - \frac{3}{4(N-2)-1}\right)\frac{\overline{Y}_S - \overline{Y}_F}{s}$ is an estimate of δ,

$w_i = \frac{1}{\hat{\sigma}^2(d_i)} \Big/ \sum_j \frac{1}{\hat{\sigma}^2(d_j)}$, and $\hat{\sigma}^2(d_i) = \frac{N}{n_F n_S} + \frac{d_i^2}{2N}$.

F2: $\hat{\sigma}(d_+) = \left(\sum_i \frac{1}{\hat{\sigma}^2(d_i)}\right)^{-1/2}$ and the 95% confidence interval is defined by $d_+ \pm 1.96 \times \hat{\sigma}(d_+)$.

F3: $\hat{\sigma}(d_{+,adj}) = \sqrt{\hat{\sigma}^2(d_{+,measure}) + \hat{\sigma}^2(d_{+,SAT})}$

F4: $2r_{AB}/(1+r_{AB})$, where r_{AB} is the correlation between Sample A and Sample B (treating the Sample A and Sample B means as parallel measurements and treating a school's mean score as a composite of the Sample A and Sample B means).

APPENDIX G: LOWEST AND HIGHEST AVAILABLE CORRELATIONS

Table G1.
Lowest and highest available student-level correlations

Construct(s)	Test	Lowest available	Second lowest	Third lowest	Third highest	Second highest	Highest available
Critical Thinking	1. MAPP	CAAP Ess. (.34)	CLA MA (.45)	CAAP Math (.46)	MAPP Writ. (.76)	CAAP Read. (.76)	MAPP Read. (.86)
	2. CAAP	CAAP Ess. (.32)	CLA MA (.39)	CLA CA (.47)	MAPP Read. (.71)	CAAP Sci. (.74)	MAPP CT (.75)
	3. CLA PT	CAAP Ess. (.32)	CAAP Math (.40)	MAPP Math (.46)	MAPP CT (.53)	MAPP Read. (.55)	CAAP CT (.58)
	4. CLA CA	CAAP Ess. (.40)	CAAP Math (.44)	MAPP Math (.46)	CAAP Read. (.50)	CAAP Sci. (.50)	MAPP CT (.52)
Writing	5. MAPP	CAAP Ess. (.33)	CLA MA (.44)	CLA CA (.48)	CAAP Writ. (.72)	MAPP Read. (.73)	MAPP CT (.76)
	6. CLA MA	CAAP Ess. (.37)	CAAP Sci. (.39)	CAAP CT (.39)	MAPP CT (.45)	CAAP Read. (.46)	CLA CA (.47)
	7. CAAP	CLA MA (.44)	CAAP Math (.48)	CLA PT (.49)	MAPP Read. (.70)	CAAP Read. (.71)	MAPP Writ. (.72)
	8. CAAP Ess.	CAAP Sci. (.28)	MAPP Math (.29)	MAPP Read. (.31)	MAPP CT (.34)	CLA MA (.37)	CLA CA (.40)
Mathematics	9. MAPP	CAAP Ess. (.29)	CLA MA (.40)	CLA CA (.46)	MAPP CT (.63)	CAAP Sci. (.71)	CAAP Math (.76)
	10. CAAP	CLA MA (.39)	CLA PT (.40)	CLA CA (.44)	CAAP Writ. (.48)	MAPP Writ. (.51)	MAPP Math (.76)
Reading	11. MAPP	CAAP Ess. (.31)	CLA MA (.43)	CAAP Math (.46)	MAPP Writ. (.73)	CAAP Read. (.76)	MAPP CT (.86)
	12. CAAP	CAAP Math (.44)	CLA MA (.46)	CLA CA (.50)	CAAP Writ. (.71)	MAPP CT (.76)	MAPP Read. (.76)
Science	13. CAAP	CAAP Ess. (.28)	CLA MA (.39)	CLA CA (.50)	MAPP Math (.71)	MAPP CT (.74)	CAAP CT (.74)

Table G2.
Lowest and highest available school-level correlations

Construct(s)	Test	Lowest available	Second lowest	Third lowest	Third highest	Second highest	Highest available
Critical Thinking	1. MAPP	CAAP Ess. (.73)	CLA MA (.88)	CAAP Math (.88)	CLA CA (.95)	MAPP Read. (.99)	MAPP Writ. (.99)
	2. CAAP	CAAP Ess. (.65)	CAAP Read. (.76)	CAAP Math (.77) CLA MA (.77)	MAPP Writ. (.92)	MAPP CT (.94)	CAAP Sci. (.95)
	3. CLA PT	CAAP Ess. (.76)	CAAP CT (.81)	CAAP Writ. (.81)	MAPP Writ. (.90)	MAPP Math (.91)	MAPP Read. (.93)
	4. CLA CA	CAAP Math (.77)	CAAP Ess. (.80)	CAAP Sci. (.83)	MAPP Read. (.95)	MAPP CT (.95)	CAAP Writ. (.96)
Writing	5. MAPP	CAAP Ess. (.75)	CLA MA (.88)	CAAP Read. (.89)	MAPP Math (.96)	MAPP Read. (.98)	MAPP CT (.99)
	6. CLA MA	CAAP Math (.77)	CAAP CT (.77) CAAP Sci. (.77)	MAPP Math (.79)	MAPP Read. (.89)	CLA CA (.91)	CAAP Read. (.91)
	7. CAAP	CAAP Ess. (.78)	CAAP Math (.79)	CLA PT (.81)	MAPP Read. (.93)	MAPP CT (.95)	CLA CA (.96) MAPP Writ. (.96)
	8. CAAP Ess.	CAAP Math (.53)	CAAP Sci. (.61)	MAPP Math (.63)	CLA CA (.80)	CLA MA (.85)	CAAP Read. (.86)
Mathematics	9. MAPP	CAAP Ess. (.63)	CAAP Read. (.79)	CLA MA (.79)	CAAP Math (.95)	MAPP Writ. (.96)	CAAP Sci. (.97)
	10. CAAP	CAAP Ess. (.53)	CAAP Read. (.74)	CLA MA (.77)	CAAP Sci. (.88) MAPP CT (.88)	MAPP Writ. (.90)	MAPP Math (.95)
Reading	11. MAPP	CAAP Ess. (.78)	CAAP Math (.86)	CLA MA (.89)	CLA CA (.95)	MAPP Writ. (.98)	MAPP CT (.99)
	12. CAAP	CAAP Math (.74)	CAAP CT (.76)	CAAP Sci. (.77)	MAPP Read. (.91)	CLA MA (.91)	CLA CA (.95)
Science	13. CAAP	CAAP Ess. (.61)	CAAP Read. (.77)	CLA MA (.77)	MAPP CT (.94)	CAAP CT (.95)	MAPP Math (.97)