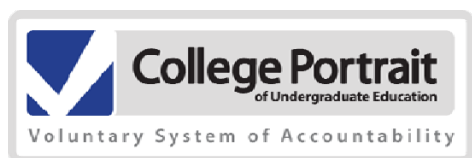


Interpretation of Findings of the Test Validity Study Conducted for the Voluntary System of Accountability

David Shulenburger, Ph.D.
Vice President, Academic Affairs
Association of Public and Land-grant Universities

Christine Keller, Ph.D.
Research Director,
Association of Public and Land-grant Universities
and
Executive Director,
Voluntary System of Accountability

November 2009



1307 New York Avenue, NW, Suite 400, Washington, DC 20005

Table of Contents

Introduction	2
Background	2
Key Findings for the VSA.....	4
Key Points for VSA Participants	10
Cautions for VSA Participants	11
Sources	11

I. INTRODUCTION

This document reflects the Voluntary System of Accountability's (VSA) perspective on how the test validity study (TVS) findings inform learning outcomes measurement within the VSA. The TVS report presents findings at both the institutional level and the individual student level; this abstract focuses primarily on institutional level findings except where reference to student level findings are necessary to fully understand institutional level results. It is not intended to be a summary of the TVS report. The TVS report and TVS executive summary are found on the VSA website at <http://www.voluntarysystem.org/index.cfm?page=research>.

In this abstract, four questions of potential concern to VSA participants are posed and relevant findings from the TVS report are reported under each question. (The TVS research questions¹ are broader than the ones we examine here and are listed in the footnote). Excerpts from the TVS report are generally quoted verbatim and are printed in italics with the page reference following in parentheses so that the reader can easily place the material quoted in the body of the TVS report. We have placed in bold italics phrases from the TVS report which most directly bear on the question under discussion. We stress that the interpretations of TVS findings contained in this document are those of the authors.

II. BACKGROUND

Two taskforces² of higher education leaders from a variety of backgrounds thoroughly evaluated sixteen potential learning outcomes tests and recommended three for options within the VSA. The VSA presidential advisory board carefully reviewed and ultimately confirmed the taskforce recommendation. Multiple test options were identified for use in VSA because public universities expressed strong desire to have the ability to select a test best suited to their particular campus circumstances. The three tests chosen were:

- **Collegiate Assessment of Academic Proficiency (CAAP)** – two modules: critical thinking and writing essay. CAAP is a product of ACT.
- **Collegiate Learning Assessment (CLA)** – complete test including a performance task and an analytic writing task (consisting of a make-an-argument and a critique-an-argument prompt). The CLA measures critical thinking, analytic reasoning, problem-solving, and written communication. CLA is a product of the Council for Aid to Education (CAE).
- **Measure of Academic Proficiency and Progress (MAPP)** – two sub scores of the test: critical thinking and written communication. MAPP is a product of Educational Testing Service (ETS).

¹ The research questions in the TVS study are: 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?* 2. *Is the difference in average scores between freshmen and seniors related to the construct tested, response format, or the test's publisher?* 3. *What are the reliabilities of school-level scores on different tests of college learning?*

² For a full description of the committee process and membership see <http://www.voluntarysystem.org/index.cfm?page=background> and for the full report of the committee see <http://www.voluntarysystem.org/docs/cp/LearningOutcomesInfo.pdf>

The taskforces determined that the CAAP, CLA, and MAPP were valid tests for measurement of critical thinking and written communication³. Two types of validity need to be distinguished: face validity and construct validity. The VSA taskforce concluded that the portions of the three tests selected for use in VSA had face validity. In other words, each of the tests presents the test taker with tasks that clearly require the use of critical thinking and written communication abilities. Face validity is very important as those considering the results must be confident that the skills being measured are those relevant and valued by future employers. However, the VSA taskforces recommended additional research to evaluate the concurrent validity across the three tests so the VSA could more confidently state that learning outcomes results were generally comparable for each of the different test options.

In fall 2007 the Fund for the Improvement of Postsecondary Education (FIPSE) funded a test validity study of the three tests of critical thinking and written communication used to measure value-added student learning outcomes within the VSA. The Association of Public and Land-grant Universities (APLU), under the direction of FIPSE grant co-principal investigator David Shulenburg, subcontracted for the testing and analytical work to be done by a consortium of testing experts led by Stephen Klein from the Council for Aid to Education (CAE), Ou Lydia Liu from Educational Testing Service (ETS), and James Scoring from ACT.

Test Frame

In fall 2008 and spring 2009 thirteen tests were administered to approximately 1100 students at thirteen⁴ colleges and universities across the U.S.⁵ The tests included the portions of CAAP, MAPP, and CLA used in VSA⁶ along with additional component tests of CAAP and MAPP: two tests in reading, two tests in mathematics and one in science. The tests and constructs are outlined in Table 1 reproduced from the full TVS report (Klein, Liu, Scoring, Bolus, Bridgeman, Kugelmass, Nemeth, Robbins, and Steedle, 2009).

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.

³ Analytic reasoning is sometimes listed as a third core skill but there is disagreement as to whether this ability is actually integral to the other two core skills so this document simply refers to two core skills.

⁴The 13 universities and colleges are Alabama A & M University, Arizona State University at the Tempe Campus, Boise State University, California State University, Northridge, Florida State University, Trinity College, Massachusetts Institute of Technology, 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.

⁵ The results of the test administration at each university are confidential and the results will not be presented in any way that serves to identify a specific university's results.

⁶ The MAPP writing essay test that is a component of VSA was not administered to students because of the great similarity of it with the CAAP writing essay. This economy was needed in order to enable the full array of tests of different constructs to be included.

Each of the 13 institutions were recruited a sample of 46 first-time, full-time freshmen and 46 seniors who had entered the institution as freshmen to take the test.⁷ Student participants were given a \$150 Amazon.com gift certificate if they completed three separate testing sessions.

III. Key Findings for the VSA

I. What is the reliability of school-level scores of different measures of writing and critical thinking ability?

Overall, the reliability of school level scores was high across all the measures of writing and critical thinking abilities. The TVS report explains the implications of high reliability at the school level in two different sections.

*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.** (Klein, et al., 2009, p. 4)*

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. (Klein, et al., 2009, pp.28-29).

Table 5 (Klein, et al., 2009, p. 29) from the TVS report details the specific reliability coefficients and is reproduced on the following page.

⁷ 1051 students took all three tests, 23 took only two tests and 51 took only one test. 51% of the students taking all three tests were freshmen and 49% were seniors, a near perfect distribution. The resulting samples were reasonable reflectors of their school's populations. Appendix C of the TVS Report has a full description of the sample and school characteristics.

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

The TVS report’s observation regarding sample size requires clarification for institutions participating in the VSA. As part of the VSA guidelines for administering one of the learning outcomes tests, participants are instructed to follow the recommendations of the appropriate testing company. At a minimum, CLA users are advised to test a minimum sample size of 100 each for freshmen and seniors; MAPP and CAAP users are advised to test a minimum of 200 each for freshmen and seniors. All three test companies recommend larger samples when a school wants to disaggregate the results by student groups. Thus the high correlations and reliable results obtained in the TVS study with samples of 30 to 50 students are useful for purposes of validation but VSA schools should continue to follow established minimums of 100 or 200 for their value-added measurement.

II. To what degree do different measures designed to assess the same construct (such as critical thinking) correlate with each other as compared to tests that are designed to assess other constructs (such as reading)?

It would be exceedingly unusual to find a test that measures only a single, unique ability. For example, essay writing and critical thinking skills are clearly intertwined. Science and math tests draw on critical thinking skills as do tests of reading comprehension. Math “word-problems” require a certain level of reading comprehension skills as well as mathematical skills. As the researchers in the TVS study state “*it is recognized that a single test may measure multiple constructs and that constructs may overlap.*” (Klein, et al., 2009, p. 11) In addition, individuals who are proficient in one domain may be proficient in another domain. For these reasons test scores generally exhibit a significant level of covariance (i.e., the test scores move in tandem). The TVS researchers describe the complexity of interpreting correlations among constructs in the following excerpt.

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 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. (Klein, et al., 2009, p. 20)

The basic correlation matrices in the TVS tables 2a and 2b (Klein, et al., 2009, p. 24) are reproduced below. Both the student- and school-level results are shown because the student-level data informs the conclusions concerning the school-level results.

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

As demonstrated in Table 2a above, the correlation patterns for the student-level results generally supported the construct validity among the different measures. As detailed in the TVS report:

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). (Klein, et al., 2009, p. 24)

... 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. (Klein, et al., 2009, p.30)

The correlation patterns at the school level (Table 2a) were parallel to the patterns at the student level but less distinct. The TVS report explains this finding in more detail.

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. (Klein, et al., 2009, p. 31)

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 ($r = .92$). (Klein, et al., 2009, p. 31)

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 ($r = .83$). (Klein, et al., 2009, p. 31)

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 ($r = .83$). (Klein, et al., 2009, p. 31)

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. (Klein, et al., 2009, p. 31)

III. Is the average difference in mean scores (effect sizes) between freshmen and seniors similar across the different measures of the same construct?

In order to compare changes in mean scores across tests with dissimilar score distributions and to control for differences in average student ACT or SAT scores, the researchers created a standardized index of “effect size.” The effect size reflects the average difference between freshmen and seniors on the TVS tests. Larger effect sizes indicate greater differences in mean scores between freshmen and seniors.

The test validity study found the average difference in mean scores between freshmen and seniors to be nearly identical across different measures of the same construct.

*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.** (Klein, et al., 2009, p. 4)*

The TVS analyses include both observed and adjusted effect sizes. An adjustment was necessary because on average seniors had higher ACT or SAT scores than freshmen. Adjusting the effect sizes created a standardized measure that could be interpreted to reflect learning gains during college rather than prior academic achievement.

The observed (or unadjusted) effect size results are described in more detail below and shown in Table 4a (Klein, et al., 2009).

*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. (Klein, et al., 2009, p. 27)*

The adjusted effect size results are described in the paragraph below.

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. (Klein, et al., 2009, p. 27)

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

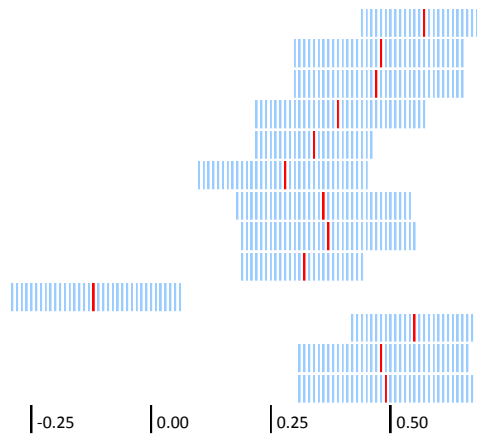


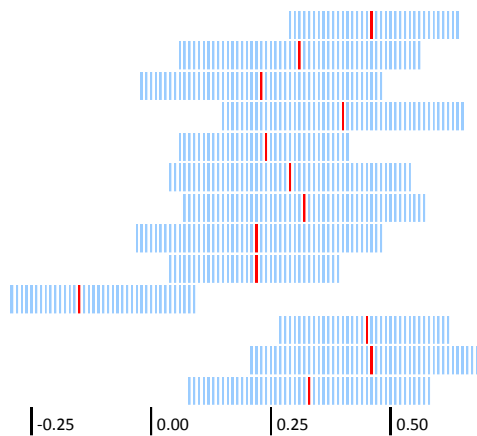
Table 4b.

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



IV. Do the scores on tests that use different response modes (such as essay versus multiple choice) to assess a given competency (such as writing ability) correlate higher with each other than they do with scores on tests that use the same response mode but assess different constructs? In other words, to what extent are the correlations among tests a function of mastery of the constructs being measured and the response modes of the tests?

The relative consistency in effect size across the three tests provide evidence that differences in score gains are associated with learning differences and not with the test or test format. More specifically:

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. (Klein, et al., 2009, p.32)

IV. KEY POINTS FOR VSA PARTICIPANTS

The TVS findings provide evidence that 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.

The 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 TVS findings allow leaders at VSA institutions to select the instrument that best fits the circumstances at their particular institution with confidence in the technical and measurement abilities of all three options. Other important considerations are described by the TVS authors.

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 systematically, 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 (Klein, et al., 2009, p. 33).

V. CAUTIONS FOR VSA PARTICIPANTS

I. The findings of the TVS study demonstrate that the three tests used within the VSA have highly correlated average scores at the school-level. The correlations are more varied and generally lower at the student-level. In particular, scores from brief, open-ended tests are less reliable at the student level.

II. Despite the high correlations among the tests measuring the same construct, especially critical thinking, the study does not “prove” that the tests measure the same thing. What the study shows is that students who do well on one test of “critical thinking” generally do well on another test of “critical thinking.”

III. Although on average, the tests provide similar adjusted effect sizes (which could be considered a measure of value-added) the TVS did not have adequate data to directly evaluate the comparability of value-added scores. The appropriate conclusion is that each of the three tests provides similar results for ordering schools by their mean test scores.

VI. SOURCES

1. Klein, S., Liu, O.L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., Nemeth, A., Robbins, S., & Steedle, J. (September 29, 2009). *Test Validity Study (TVS) Report*. Supported by the Fund for Improvement of Postsecondary Education (FIPSE). Online at: <http://www.voluntarysystem.org/index.cfm?page=research>.