Comparison of the VSA Learning Outcomes Measures

Results from a Test Validity Study

Christine Keller, APLU

Heather Kugelmass, CAE
Alex Nemeth, CAE

CAIR Annual Conference
Sacramento, California
November 18-20, 2009

Voluntary System of Accountability (VSA)

Initiative by public 4-year universities to supply basic, comparable information on the undergraduate student experience to important constituencies through a common web report – the College Portrait.

Sponsored by APLU & AASCU. Funding from Lumina.

VSA Goals

- Provide a useful tool for students during the college search process
- Assemble information that is transparent, comparable, and understandable
- Demonstrate accountability and stewardship to public
- Measure educational outcomes to identify and enhance effective educational practices

VSA Context

- Increasing disinvestment in higher education
- Policy-makers and employers want evidence of educational outcomes – particularly broad transferable skills
- Perceived lack of useful and transparent data prevents institutions from demonstrating accountability and contribution to public good (*Spellings Commission*)
- Better for higher education to tackle the challenges of measuring learning outcomes than to have it imposed from the outside

Background: SLO in VSA

Measure student improvement (learning gains) in critical thinking, analytical reasoning, problem solving, and written communication at the institution level

- 16 tests evaluated by 2 task forces, 3 selected
  - **CAAP**: Collegiate Assessment of Academic Proficiency - ACT
  - **CLA**: Collegiate Learning Assessment - CAE
  - **MAPP**: Measure of Academic Proficiency and Progress - ETS

SLO: VSA Reporting

- Learning gains (value-added) - the difference between actual and expected scores of graduating and entering students after controlling for academic ability
  - CAAP – 2 modules: critical thinking, written communication
  - CLA – complete test including performance tasks, analytic writing tasks
  - MAPP – 2 test subscores: critical thinking, written communication

Additional Research Needed

VSA needed evidence of relationships among CAAP, CLA, MAPP

- Fund for the Improvement of Post-Secondary Education (FIPSE) provided funding for a Test Validity Study
- Allowed systematic evaluation of results when students across different institutions took each of the three tests

Test Validity Study (TVS)

Design

Methods

Findings

Study Parameters

- 13 tests administered to 1,100 students at 13 universities
- Test components of CLA, CAAP, MAPP
  - 4 tests of critical thinking*
  - 2 tests of reading
  - 2 tests of mathematics
  - 4 tests of writing*
  - 1 test of science

**Note that only tests of critical thinking and writing are used within the VSA*

Participating Universities

- Alabama A&M
- Arizona State
- Boise State
- California State-Northridge
- Florida State
- MIT
- Trinity College
- University of Colorado-Denver
- University of Michigan, Ann Arbor
- University of Minnesota, Twin Cities
- University of Texas, El Paso
- University of Vermont
- University of Wisconsin, Stout

Sample Parameters

	Range at Institutions	Overall Mean
SAT mean scores	1000 - 1458	1181
% Minority	4% - 97%	31%
% Women	37% - 63%	55%

Sampling / Testing Methods

- 46 freshmen and 46 seniors
  - 18 years old, SAT/ACT on file
  - Part of first-time, full-time cohort
- Students asked to complete three tests
  - $150 Amazon.com / post-paid
- Standardized administration
- Counterbalancing

Research Question 1

- What are the relationships among scores on the tests?
  - 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 (ACT, CAE, ETS)?

Methods 1

- Student and school-level correlations
- Freshman and senior correlations highly similar
  - Combined to increase sample size

Findings: Correlations

- General pattern of correlations at student level support test construct validity
- Correlations very high when the school is the unit of analysis
- Mean correlation = .92 for 9 multiple choice tests
- Mean correlation = .84 for 4 constructed response measures
- Mean correlation = .85 for multiple choice tests and constructed response measures of different constructs

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?

Methods 2

- Need common scale
  - Effect sizes in standard deviation units
- School effect sizes combined
  - Precision-weighted composite effect size
- Need ability difference control
  - Adjusted effect size

Findings: Effect Sizes

- Larger effect sizes indicate greater differences in freshman and senior scores
- Seniors had higher mean scores than freshmen on all tests except the CAAP mathematics exam
- Effect sizes not systematically related to constructs, response format, or test publisher

Findings: Effect Sizes (cont'd)

- Adjusted effect sizes across 12 tests range from approximately one-quarter to one-half SD (CAAP math excluded)
- Adjusted effect sizes
  - CAAP = .33 (excluding math test)
  - CLA = .31
  - MAPP = .34

Research Question 3

- What are the reliabilities of school-level scores?

Methods 3

- Reliability calculated at the school level
- Modified split-sample approach
  - Students split randomly into sample A and B
  - Mean scores for sample A and B
  - Correlations of mean scores across schools
  - Repeated 1,000 times
  - Spearman-Brown correction for sample size
  - Adjusted reliabilities reported by class

Findings: Reliability

- Reliability is score consistency
- When the school was the unit of analysis, across the 13 tests:
  - mean reliability = .87
  - lowest reliability = .75
- Conclusion: Score reliability is not a concern

Summary of Findings

- Across constructs, response formats, and test publishers
  - School-level correlations high
  - Effect sizes consistent
  - School-level reliabilities high

Overall Conclusions

- CAAP, CLA, MAPP provide similar results for ordering schools by mean scores
- All tests rank schools similarly, regardless of the construct, response format, or publisher
  - *TVS did not have adequate data to directly test comparability of value-added scores*
- Students who do well on one test of critical thinking generally do well on another test of critical thinking
  - *High correlations do not "prove" the tests measure the same construct*

Implications for VSA

- VSA institutions continue to select from CAAP, CLA, or MAPP to administer and report
- Technical and measurement abilities consistent across tests
- Important considerations for selection:
  - Acceptance by students, faculty, administrators or other policy makers
  - Trade-offs in cost, ease of administration, etc.
  - Utility of the test for other purposes - supporting campus activities and services or providing guidance on improving learning

TVS Reports

- 3 reports are available on the VSA website http://www.voluntarysystem.org/index.cfm
  - The complete TVS report
  - TVS Executive Summary
  - An interpretative summary by VSA, especially for VSA participating schools

More Information

Christine Keller
ckeller@aplu.org
APLU Director of Research & Policy Analysis
VSA Executive Director

Extra Slides
(tables from TVS Report)

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

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

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