



cla+

Reliability and Validity of CLA+

Doris Zahner
Director of Test Development & Measurement Science

The skills needed by today's students are far greater than those required of students just a decade ago. A greater emphasis is being given to college readiness and college outcomes because employers now, more than ever, expect college graduates to possess writing, critical thinking, and problem solving skills (Hart Research Associates, 2006) in response to the changing demands of available jobs (Autor, Levy, & Murnane, 2003). Students can no longer rely solely on the accumulation of disciplinary knowledge and skills. The educational community has begun to emphasize so-called "21st century" skills (PARCC, 2012; SBAC, 2012) in addition to knowledge in specific content domains (Arum & Roksa, 2011; Porter, McMaken, Hwang, & Yang, 2011; Silva, 2008; Wagner, 2008) in hopes of fostering the development of critical thinking, problem solving, communication, collaboration, creativity and innovation skills (Porter, et al., 2011). In fact, nearly 80% of Association of American Colleges and Universities member institutions have a list of general learning outcomes intended for all students regardless of their academic programs, and skills such as critical thinking and writing are among the most commonly included (Hart Research Associates, 2009).

Since 2002, the Council for Aid to Education (CAE) has pioneered the use of performance-based assessments for determining whether students can successfully analyze a body of information and communicate that information in an open-ended response. To date over 700 institutions, both in the United States and internationally, have participated in our performance assessments, either through our flagship product, the Collegiate Learning Assessment (CLA) or its sister assessment, the College and Work Readiness Assessment (CWRA).

Performance assessments are open-ended instruments that require students to demonstrate their knowledge, skills, and abilities by generating their own solutions and responses to a given problem rather than selecting the correct answer from a given list. This type of assessment is directly aligned with current national reform efforts in the K-12 arena, which are aimed at improving teaching and learning (NGA & CCSSO, 2010a, 2010b; Partnership for 21st Century Skills, 2012). The CLA was originally designed to measure an institution's contribution, or value-added, to the development of the higher-order thinking skills of its students (Klein & Benjamin, 2008; Klein, Benjamin, Shavelson, & Bolus, 2007). Therefore the institution, not the individual student, was the primary unit of analysis. This approach allowed institutions to compare their improvements on the CLA with those at similarly selective institutions, and use that information to improve teaching and learning.

Ten years later, CAE is launching a new and improved version of the CLA, the CLA+. The CLA+ retains aspects of the CLA that have made it novel and indispensable for educational improvement. Chief among these is the Performance Task (PT). The original CLA PTs assessed four components of higher-order skills: Analytic Reasoning and Evaluation, Problem Solving, Writing Effectiveness, and Writing Mechanics. The CLA+ PTs are improved versions of this original concept and remain the anchor of the assessment. The new PTs measure similar constructs: Analysis and Problem Solving, Writing Effectiveness, and Writing Mechanics. However, new Selected Response Items (SRIs), which measure analysis and problem solving skills, are now being introduced. These selected-response items, like the PTs, are all anchored to documents that emulate real-world scenarios or problems. They are far from the typical recall and recognition multiple-choice items seen in many other standardized assessments.

CAE decided to include selected-response items in the CLA+ in order to improve the precision of student-level results. This report provides an overview and results of the pilot study for the CLA+.

The assessment approach and the structure of the test distinguish the CLA+ from other assessments of critical thinking. During the first 60 minutes of the test, students complete an integrated performance task that mirrors a real-world challenges that could be encountered in a work or academic environment. The student is provided with three or more documents such as a data table or graph, a newspaper article, a research report, or other critical information sources that students typically encounter in real-world settings. Students are asked to critically read and analyze the information presented in the documents and then generate a written response. Typically, students are asked to make a decision about the scenario presented in the PT, provide supporting evidence from the documents, and refute the opposing argument.

The final 30 minutes of the CLA+ consists of a set of 25 selected-response items (SRIs). These selected-response items are aligned to the same critical thinking skills assessed by the PTs. This section of the examination assesses three different constructs: scientific and quantitative reasoning (10 items), critical reading and evaluation (10 items), and critiquing an argument (5 items). Like the Performance Tasks, each problem set is anchored to authentic source documents that require careful analysis and evaluation of information.

The reliability and validity data reported in this paper pertains to a pilot study of the CLA+ conducted during the spring of 2012 with four participating higher-education institutions. The version of the CLA+ used for the pilot study was longer than the current CLA+. It consisted of two 50-minute PTs and a set of 30 SRIs. Participants in the pilot study had a total of 150 minutes to complete the assessment. Based on feedback from the participating institutions and current CLA clients, CAE shortened the CLA+ from 150 minutes to 90 minutes to be consistent with current the CLA test time. The operational CLA+ will consist of one 60-minute PT and 25 SRIs to be completed in 30 minutes.

While the results presented in this paper pertain specifically to data from the CLA+ pilot study and the traditional CLA, the operational CLA+ uses versions of the CLA and CLA+ pilot study PTs and SRIs from the pilot study. We are confident that the reliability and validity of the operational CLA+ will be similar to the longer pilot study version of the CLA+. Since there are fewer items, reliability may be slightly lower than those reported, but sufficient individual student-level reliability will still be achievable.

RELIABILITY

Traditionally, CLA scores have been very reliable at the institution level ($\alpha = .80$) (Klein, et al., 2007), but not at the individual student level (alternate forms reliability = .45). This is due to the fact that, at the individual student level, the CLA was only a single PT or Analytic Writing Task (Make-an-Argument and Critique-an-Argument). Reliability was achieved only when CLA scores were aggregated across all students at a participating institution. The CLA+ has students answering more than one or two questions, and these questions all focus on the same constructs (e.g., analysis and problem solving), resulting in higher reliability of individual student scores.

CLA+ Pilot Study

Cronbach's alpha, which ranges from 0 to 1, measures internal consistency reliability of a set of items (i.e., whether the items on an assessment are positively correlated). Table 1 shows the reliability coefficients for different sets of items for the two forms of the CLA+ administered in the pilot study.

Table 1: Cronbach's alpha statistics for CLA+ items

| | Form A | Form B |
|------------------------|--------|--------|
| Total (2PTs + 30 SRIs) | .87 | .85 |
| PTs only | .43 | .57 |
| SRIs only | .80 | .78 |
| SRI_SQR | .43 | .62 |
| SRI_CRE | .58 | .44 |
| SRI_CA | .67 | .60 |

Using the inter-rater correlations as the PT reliability on each of two PTs, reliability was between .67 and .75 across the four PTs. Reliability for the SRIs ($\alpha = .80$ and .78) is higher than the PTs. Using a stratified alpha (Cronbach, Schonemann, & McKie, 1965) for combining the PTs with the SRIs, we obtain a reliability coefficient of .87 for Form A and .85 for Form B. These coefficients reflect the combination of 5 subtests: SRI-SQR, SRI-CRE, SRI-CA, PT-SQR and PT-CRE. With the addition of the SRIs to the PTs, individual student scores are now reliable enough for interpretation at the individual level and for making decisions about grading, scholarships, admission, or placement.

STUDENT EFFORT AND ENGAGEMENT SURVEY RESPONSES

Tables 2 – 5 summarize results for some of the questions from the student survey administered to CLA + pilot study participants following the completion of the assessment.

Testing Time

Results in Table 2 show that the large majority (at least 90%) of students finished the different sections of the test with some time remaining. This was particularly true for the 45 minutes allotted to the

SRI section of the test, where less than 1% of the participants did not have sufficient time to complete the test (Table 2).

Table 2: Allotted Time

Did you have enough time to finish:

| | PT_SQR | PT_CRE | SRI |
|---|--------|--------|-------|
| I finished the task with plenty of time remaining. | 67.4% | 79.4% | 84.8% |
| I finished the task with little time remaining. | 22.0% | 16.2% | 13.2% |
| I finished the task with no time remaining. | 4.4% | 3.2% | .9% |
| I did not finish the task. I would need a little more time to finish. | 5.6% | 1.2% | .9% |
| I did not finish the task. I would need a lot more time to finish | .6% | 0% | .3% |

Effort

Tables 3 and 4 show that most students put at least a moderate amount of effort into their CLA+ responses (94%) and found the tasks to be at least moderately engaging (79.7%). These results are encouraging because low student motivation and effort are threats to the validity of test score interpretations. If students are not motivated, their scores will not be accurate reflections of their true competency levels. Although studies of the CLA have shown that aggregate student motivation is not a significant predictor of aggregate CLA performance (Steedle, 2010), this is not the case for results at the individual student level. Research suggests that adding stakes to an assessment increases student motivation and student performance (Liu, Bridgeman, & Adler, 2012). Although the data collected in the survey was self-reported, the new CLA+ may have stakes attached, which should improve the validity of CLA+ score interpretations.

Table 3: Effort

How much effort did you put into these tasks?

| | |
|-----------------------------|-------|
| No effort at all | 0.6% |
| A little effort | 5.0% |
| A moderate amount of effort | 44.9% |
| A lot of effort | 34.3% |
| My best effort | 15.2% |

Table 4: Engagement

How engaging did you find the tasks?

| | |
|---------------------|-------|
| Not at all engaging | 3.2% |
| Slightly engaging | 17.0% |
| Moderately engaging | 51.6% |
| Very engaging | 24.6% |
| Extremely engaging | 3.5% |

VALIDITY

Construct Validity

In the fall semester of 2008, CAE collaborated in a construct validity study with ACT and ETS to investigate the construct validity of three assessments: the CLA, the Collegiate Assessment of Academic Proficiency, and the Measure of Academic Proficiency and Progress (Klein et al., 2009). Construct validity indicates the degree to which test scores may be interpreted as indicators of the particular skill (or construct) that the test purports to measure. Construct validity is often evaluated by examining by the pattern of correlations between a test and other tests of similar and different skills (Campbell, 1959). For example, if the CLA measures critical thinking skills, then it should be highly (positively) correlated with other tasks that measure critical thinking.

Results from the study show that for critical thinking, the CLA is indeed strongly positively correlated with other tasks that measure critical thinking. The correlation between CLA Performance Tasks and other tests of critical thinking range from .73 to .83. The correlation between CLA Critique-an-Argument tasks and other constructs that measure critical thinking range from .73 to .93. A full report of the Test Validity Study (Klein, et al., 2009) can be found on CAE's website at http://www.cae.org/content/pdf/TVS_Report.pdf.

Face Validity

A test is said to have face validity when, on the surface, it appears to measure what it claims to measure. For the CLA+ to have face validity, the tasks must emulate the critical thinking and writing challenges that students will face both in and out of the classroom. In the CLA+ pilot study, students were asked how well they thought the tasks on the CLA+ measured writing, reading comprehension, mathematics, and critical thinking and problem solving skills.

As shown in Table 5, results indicate that students perceived the tasks to be at least moderately good assessments of writing (78.5%), reading comprehension (88.0%), and critical thinking and problem solving (86.2%) skills. The scientific and quantitative reasoning sections of the CLA+ measure reasoning skills such as identifying whether quantitative data that is presented in the tasks are connected or conflicting or making inferences based on data that are presented to them. The focus is more on quantitative literacy rather than traditional mathematical problem solving skills. For example, student may be expected to recognize that a correlation between two variables does not necessarily imply a causal relationship. So it appears that we are measuring what we purport to measure on the CLA+ tasks.

Table 5: Face Validity

How well do you think these tasks measure the following skills:

| | Writing | Reading Comprehension | Mathematics | Critical Thinking & Problem Solving |
|-----------------|---------|-----------------------|-------------|-------------------------------------|
| Not well at all | 3.2% | 1.5% | 55.0% | 1.5% |
| Slightly well | 18.2% | 10.6% | 27.4% | 12.3% |
| Moderately well | 46.3% | 36.2% | 15.3% | 30.5% |
| Very well | 28.7% | 40.9% | 1.8% | 43.1% |
| Extremely well | 3.5% | 10.9% | .6% | 12.6% |

REFERENCES

- Arum, R., & Roksa, J. (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago, IL: University of Chicago Press.
- Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4), 1279-1333.
- Campbell, D. T. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.
- Cronbach, L. J., Schonemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, 25, 291-312.
- Hart Research Associates. (2006). *How Should Colleges Prepare Students to Succeed in Today's Global Economy? - Based on Surveys Among Employers and Recent College Graduates*. Washington, DC: Hart Research Associates.
- Hart Research Associates. (2009). *Learning and Assessment: Trends in Undergraduate Education - A Survey Among Members of The Association of American Colleges and Universities*. Washington, DC: Hart Research Associates.
- Klein, S., & Benjamin, R. (2008). *The Collegiate Learning Assessment (CLA)*. Council for Aid to Education. New York, NY.
- Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment: Facts and Fantasies. *Evaluation Review*, 31(5), 415-439.
- Klein, S., Liu, O. L., Scoring, J., Bolus, R., Bridgeman, B., Kugelmass, H., . . . Steedle, J. (2009). Test Validity Study (TVS) Report. Supported by the Fund for the Improvement of Postsecondary Education, from http://www.cae.org/content/pdf/TVS_Report.pdf
- Liu, O. L., Bridgeman, B., & Adler, R. (2012). *Measuring learning outcomes in higher education: Motivation matters*. Princeton, NJ: ETS.
- NGA, & CCSSO. (2010a). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects*. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers,.
- NGA, & CCSSO. (2010b). *Common Core State Standards for Mathematics*. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers,.
- PARCC. (2012). *Partnership for Assessment of Readiness for College and Careers*. Retrieved November 20, 2012, 2012, from <http://www.parcconline.org/about-parcc>
- Partnership for 21st Century Skills. (2012). *Framework for 21st Century Learning*. Retrieved November 30, 2012, 2012, from <http://www.p21.org/overview>
- Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The New US Intended Curriculum. *Educational Researcher*, 40(3), 103-116.
- SBAC. (2012). *Smarter Balanced Assessment Consortium*. Retrieved November 20, 2012, 2012, from <http://www.smarterbalanced.org/>
- Silva, E. (2008). *Measuring Skills for the 21st Century*. Washington, DC: Education Sector.
- Steedle, J. T. (2010). *Incentives, Motivation, and Performance on a Low-Stakes Test of College Learning*. Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.
- Wagner, T. (2008). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need--and what we can do about it*. New York, NY: Basic Books.