

# CAE

## Three Principle Questions About Critical-Thinking Tests<sup>1</sup>

Dr. Roger Benjamin  
President of the Council for Aid to Education



There are three important questions regarding what the Collegiate Learning Assessment (CLA) measures and why it is important. First, what is the rationale for assessing critical-thinking skills<sup>2</sup> independently as compared to academic disciplines? Second, are there unavoidable interaction effects between measures of critical-thinking skills and academic disciplines? In other words, do students in some disciplines do better on performance assessments (the constructed-response tests the CLA uses to measure critical-thinking skills) than others? Third, what evidence is there that these skills can be measured and are of practical use? Below are CAE's answers to these questions.

## The Rationale for Critical Thinking Skills

Critical-thinking skills—defined as analytic reasoning, quantitative reasoning, problem solving, and writing—fill an important gap not dealt with by academic majors. Students major in disciplines that faculty are organized within and support. It is perhaps natural that faculty view their respective disciplines as the core products of undergraduate education. However, in recent decades interest in improving critical-thinking skills has increased significantly. Definitions of knowledge and learning increasingly focus on the ability to apply learned skills to new situations.<sup>3</sup> In today's global Knowledge Economy the ability to access, structure, and use information becomes more essential than merely having command of specific discipline-based content. Thus, a central focus of undergraduate education is teaching and improving critical-thinking skills both independently and within disciplines.

If there is merit to focusing on critical-thinking-skills, examining them only within the context of disciplines of inquiry or through analytic constructs such as the humanities, natural sciences, physical sciences, or social sciences, commits what statisticians call the individualistic fallacy<sup>4</sup>—the parts do not add up to define undergraduate education as a whole for students. This is because the development of these core cognitive skills is a joint product of the courses and experiences students encounter over their four years of undergraduate study. Moreover, because there is a holistic quality about these cognitive skills, it is important to assess them with measurement instruments that are able to capture their holistic quality. Performance assessments are able to carry out this task. Multiple-choice tests, alone, do not.

## Potential Interaction Effects Between Critical-Thinking Skills and Academic Major

Questions remain as to whether critical-thinking skills are independent from discipline-based skills. There are also arguments claiming that these skills cannot be measured independently from academic disciplines.<sup>5</sup> The following logic is used in CLA Performance Tasks.

As an example, consider a teacher who instructs students in her chemistry course on how to assess the characteristics of different substances, such as how they each respond to fire (the so-called “flame test”). The instructor then gives each student a different “unknown” and asks him or her to determine its chemical composition. Students are evaluated not only on their ability to figure out what the unknown substance is but also on the appropriateness of the tests they ran, the sequence of those tests, and the rationale for their decisions and conclusion.

This “unknown-substance” test certainly requires substantive and procedural knowledge of chemistry (such as how to run a flame test) but it also assesses generic problem-solving and reasoning skills. So a task that provides students with the knowledge they need (in the figures, graphs, technical reports, and newspaper articles in the “Document Library”) can focus on assessing critical thinking. That is what the CLA does and why there is no empirical interaction between the substantive context/setting for a performance-task prompt and an examinee's academic major. The performance-task format, structure, and approach do a good job of isolating the skills and abilities we want to measure.

Two peer-reviewed papers present corroborative evidence to support this point. In the first paper, S. Klein, et al. (2008), discuss findings by R. Shavelson on the interaction between performance-task content and

academic majors. Klein and colleagues noted how Shavelson (2010) investigated this issue using college seniors who took a CLA Performance Task during spring 2007. Each performance task was assigned to one of three content areas: science, social science, or the humanities. Students self-identified the area of their major as science and engineering, social science, humanities, or other. Ultimately, Shavelson constructed five student-level regression equations using combinations of measures of the students' entering academic ability, the SAT, and indicator variables for task area and academic major area to predict CLA scores. When SAT scores are included in the model, other variables have almost no effect on predictive accuracy. A more recent study using data from 12,632 graduating seniors from 236 four-year institutions in the United States corroborates Shavelson's findings (Steedle & Bradley, 2012). In this study, there were no significant interactions between CLA Performance Tasks and academic disciplines. This does not mean that what one studies has no effect on performance on tests of critical thinking. Overall, Steedle and Bradley (2012) and Arum and Roksa (2011) find that students who majored in disciplines in the arts and sciences, including the humanities, foreign languages, physical and natural sciences, mathematics, and engineering did better than academic majors in applied professional fields such as health, education, and business. In other words, students majoring in the arts and sciences tend to do better on all of the performance tasks than do students in applied professional fields. However, science majors do not do better on performance tasks set in a science context than they do on performance tasks set in a business context. Why might arts-and-science or engineering students do better on performance tasks overall? One hypothesis is that there is more writing and analysis required of students in those fields.

## What Evidence is there to Demonstrate that Measuring Critical-Thinking Skills Has Practical Uses for Colleges and Universities?

Let us examine whether empirical results of critical-thinking skills tests are of practical utility for the post-secondary education sector. To set the stage, many readers may be familiar with the controversial findings of *Academically Adrift* (Arum & Roksa, 2011) and the extended commentary about its findings that called into question whether college was fostering sufficient learning by students. This study was, in part, based on a sample of 25 colleges and universities that tested students longitudinally with CLA Performance Tasks over a four-year period.<sup>6</sup> If we can shed further light on this issue, it may help advance our views on this subject and other related subjects to improve student learning.<sup>7</sup>

Table 1 presents the effect-size statistics for all the colleges and universities testing with the CLA between the 2005-06 and 2011-12 academic years.<sup>8</sup> Effect size is a "simple way of quantifying the difference between two groups" (Cole, 2002). In this case, a sample of graduating seniors is compared to a sample of entering freshmen. The data are aggregated at the institutional level which is the most appropriate way to analyze CLA results because the CLA protocol relies upon a matrix sampling approach. Each student takes either a performance task or an analytic writing task (comprised of make-an-argument and critique-an-argument prompts) but not both. Therefore, the results are not reliable at the student level but are reliable at the institutional level.

The results are noteworthy for several reasons.

(1) The average effect size reflecting differences in CLA performance between entering freshmen and graduating seniors was .73 over several test administrations. The distribution of effect sizes shown in Figure 1 suggests two things: that CLA scores increase over the course of college and that some colleges contribute significantly more than others in terms of learning. The distribution of value-added scores shown in Figure 2 corroborates the latter point (note that value-added scores are set to have a mean of 0 and a standard deviation of 1).

(2) Most importantly, the symmetric bell-shaped curves of Figures 1 and 2 present a balanced picture of the collegiate landscape of student learning. Institutions with very low effect sizes or value-added

scores are in precarious territory, but there are also significant numbers of institutions with high growth measures that are clear candidates for best-practice investigations. The results suggest that it is time to implement a portfolio of studies aimed at understanding the pedagogical practices of the institutions in the high end of the bell-shaped curves.<sup>9</sup>

(3) This distribution can be used to continually gauge acquisition of critical-thinking and writing skills. Regardless of whether institutions are above or below the average effect size or value-added score, why not commit to shift the entire distribution upward by 10% over five years, then another 10%, and so on?

CLA Effect Size Statistics, All schools, 2005-2012							
Year	N	Mean	St. Dev.	Percentile Score		Range	
				25th	75th	Lower	Upper
2011-12	158	0.78	0.51	0.48	1.06	-0.76	2.30
2010-11	180	0.77	0.42	0.50	1.04	-0.23	1.84
2009-10	150	0.70	0.41	0.43	0.90	-0.34	2.21
2008-09	167	0.73	0.37	0.48	0.97	-0.09	1.92
2007-08	167	0.69	0.40	0.45	0.93	-0.39	2.11
2006-07	108	0.71	0.48	0.37	1.00	-0.58	2.05
2005-06	108	0.70	0.48	0.42	1.02	-1.02	1.77
<b>Average</b>	<b>148</b>	<b>0.73</b>	<b>0.44</b>	<b>0.44</b>	<b>0.99</b>	<b>-0.49</b>	<b>2.03</b>
<b>All Years Combined</b>	<b>103</b>	<b>0.73</b>	<b>0.44</b>	<b>0.44</b>	<b>0.99</b>	<b>-1.02</b>	<b>2.30</b>

Table 1 (shaded area corresponds with Figure 1)

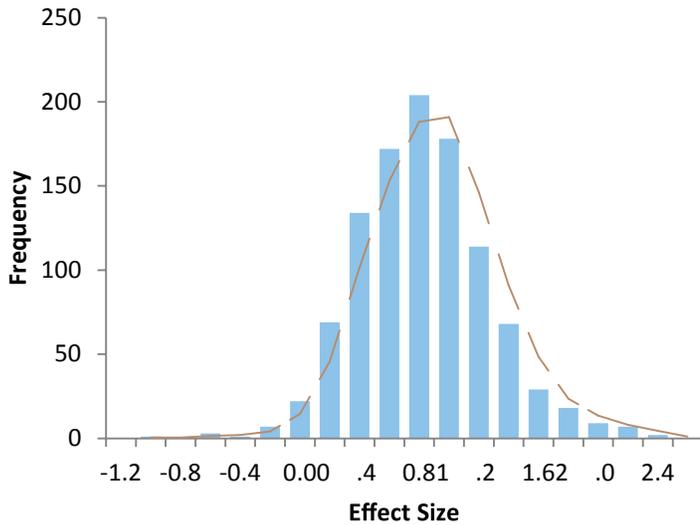


Figure 1 (histogram of effect size across all institutions testing between 2005 and 2012) (n=1038)

CLA Value-Added Score Statistics, 2009-2012							
Year	N	Mean	St. Dev.	Percentile Score		Range	
				25th	75th	Lower	Upper
2011-12	150	0.01	0.97	56	0.64	-2.88	2.58
2010-11	176	0.00	0.98	57	0.74	-2.81	2.74
2009-10	155	0.01	1.00	56	0.71	-2.75	4.34
Average	160	0.01	0.98	56	0.69	-2.81	3.22
All Years Combined	481	0.01	0.98	56	0.70	-2.88	4.34

Table 2 (shaded area corresponds with Figure 2)

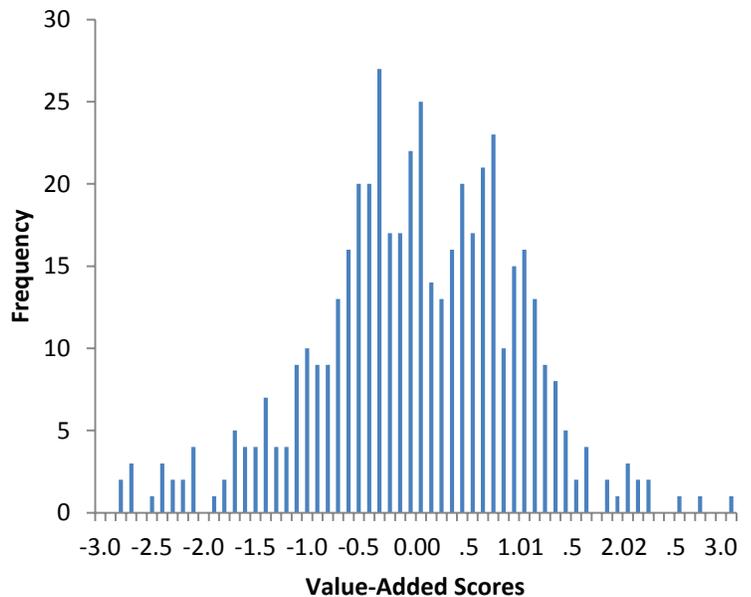


Figure 2 (histogram of value-added scores across all institutions testing between 2009 and 2012) (n=481)

## Conclusion

Entrants to today's work force can expect to change jobs several times over the course of their career. The skills required for specific jobs are likely to change quickly as well. In contrast, the importance of the core critical-thinking skills is enduring. There are additional core cognitive skills such as leadership, teamwork, and moral decision-making worthy of measuring as soon as it can be demonstrated that they can be measured reliably. Over time, research will attempt to broaden the reach of the core cognitive skills measured. However, we have made a good start. Critical-thinking skills—at the heart of virtually all definitions of core cognitive skills—can be measured, and measured in a way that is independent from academic disciplines. The tables and figures presented here show that college matters significantly and that some colleges and universities exhibit more impact on the growth of critical-thinking skills than others. This means there is a wide canvas for faculty and administrators to study best practices to improve teaching and learning.

In conclusion, CLA results can be usefully applied to improve educational programs. The CLA is a standardized test (a test administered in the same conditions for all examinees), and it is often the belief that such assessments are not useful for improving classroom instruction. However, there is increasing evidence that performance tasks like those included in the CLA can play an important role in classroom learning and assessment (Chun, 2010). This is important because in order for faculty to take assessment seriously they must consider assessment instruments as authentic and useful to them in the classroom.

- Arum, R. & Roksa, J. (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago, Ill.: University of Chicago Press.
- Banta, T. & Pike, G. (2012). *Making the Case Against---One More Time*. National Institute for Learning Outcomes Assessment (NILOA). Occasional Paper #15, September, 24-30.
- Benjamin, R., Klein, S., Steedle, J., Zahner, D., Elliot, S., Patterson, J. (2012). *The Case for Generic Skills and Performance Assessment in the United States and International Settings*. [www.cae.org](http://www.cae.org).
- Bransford, J., Brown, A. & Cocking R. (eds.) *How People Learn*. Washington D. C. National Academy Press.
- Cole, R. (2002). "It's The Effect Size, Stupid---What effect size is and why it is important." Paper presented at the Annual Conference of the British Educational Research Association, Exeter, England: University of Exeter.
- Ewell, P. (2012). "A World of Assessment: OECD's AHELO Initiative." *Change*. September/October, 35-42.
- Keeling, R. & Hersh, R. (2011). *We're Losing Our Minds: Rethinking American Higher Education*. New York: Palgrave MacMillan.
- Klein, S., Freedman, D., & Bolus, R. (2008). "Assessing School Effectiveness." *Evaluation Review*. 32 (6), 511-525.
- Klein, S. (2009). "CLA Lumina Longitudinal Study Summary Findings." [www.cae.org](http://www.cae.org).
- NCAT.org
- NILOA.org
- Pascarella, E., Blaich, C., Hanson, J., Martin, G. (2011). "How Robust Are The Findings of Academically Adrift?" *Change*. Vol. 43, Issue 3, pp. 2024.
- Robinson, W. (1950). "Ecological Correlations and the Behavior of Individuals." *American Sociological Review*. 15: 351-357.
- Steedle, J. & Bradley, M. (2012). "Majors matter: Differential Performance on a Test of General College Outcomes." Paper presented at the Annual Meeting of the American Educational Research Association, Vancouver, Canada.

# End Notes

- (1) The U.S. version of the CLA refers to critical-thinking skills, while they are labeled generic skills in international settings such as those measured by the OECD's Assessment of Higher Education Learning Outcomes (AHELO) Initiative. (Benjamin et al., 2012; Ewell, 2012). Higher-order skills are also referred to as cognitive skills (see Benjamin, et al., 2012, for a more detailed discussion of these terms). We will use the term critical-thinking skills here.
- (2) We present a more complete rationale for generic skills and performance assessment in Benjamin, et al. (2012).
- (3) See Bransford, et al. (2000).
- (4) The classic study on this concept is Robinson (1950).
- (5) See Banta and Pike (2012) and Ewell (2012).
- (6) The average four-year effect size observed by Arum and Roksa (2011) was corroborated by a similar study that used an alternative measure of critical-thinking skills (Pascarella, Blaich, Martin, & Hanson, 2011).
- (7) In addition to the results based on the CLA presented here, two other national tests of critical thinking— the Proficiency Profile, authored by ETS and the Collegiate Assessment of Academic Progress (CAAP)—also have significant numbers of students tested to address these issues.
- (8) The tables and figures presented here are based on cross-sectional samples of entering freshmen and graduating seniors designed to simulate longitudinal studies for four years of college. For CAE studies that find the cross-sectional approach comparable to longitudinal studies see Klein, et al., (2008) and Klein, et. al., (2009).
- (9) See, for example, the studies on pedagogy and curriculum development practices published by NILOA (National Institute of Learning Outcomes Assessment); NILOA.org. National Center for Academic Transformation, NCAT.org and the argument provided by Keeling and Hersh (2011).