[collegiatelearningassessment]

FREQUENTLY ASKED TECHNICAL QUESTIONS

Council for Aid to Education
215 Lexington Ave., Floor 16
New York, NY 10016
Phone: 212.217.0700
Fax: 212.661.9766
Email: cla@cae.org
Web: www.cae.org/cla

LETTER FROM THE PRESIDENT

Dear Colleagues,

Performance-based assessments are anchored in a number of psychometric assumptions different from those employed by common multiple-choice exams. As such, initiatives like the Collegiate Learning Assessment (CLA) represent a paradigm shift, the technical underpinnings of which remain unfamiliar to many faculty and institutional researchers. The CLA is novel because it calls for 1) the institution, not the student, to be the initial unit of analysis, 2) a matrix sampling approach, and 3) a value added method which, in turn, requires evidence of the competencies students bring to college. Rightly, colleagues at campuses using or contemplating using the CLA often have a number of important questions. With this in mind, we intend to publish a technical manual that deals with, as comprehensively as possible, the many technical questions associated with the CLA. In the meantime, I hope you will be pleased with this document that addresses several of the major questions our colleagues frequently ask.

We could not continue to improve the CLA without your constructive advice. We would appreciate receiving any new questions that these responses suggest to you.

Sincerely,

Roger Benjamin, Ph.D.
President & CEO, Council for Aid to Education

TABLE OF CONTENTS

INTRODUCTION

What is the CLA?

The Collegiate Learning Assessment (CLA) is a major initiative of the Council for Aid to Education. Along with the Community College Learning Assessment (CCLA) and the secondary-education-level College and Work Readiness Assessment (CWRA), the CLA offers a constructed-response approach to the assessment of higher-order skills. Although the assessment for each level of institution (high school, university, and community college) has a unique name, any mention of the CLA in this document will pertain to all three tests, unless otherwise indicated.

The CLA is designed to measure an institution's contribution, or value added, to the development of higher-order skills, and therefore the institution—not the student—is the primary unit of analysis. This approach allows an institution to compare its student learning results on the CLA with the learning results at similarly selective institutions, and use that information to improve teaching and learning.

The CLA uses constructed-response tasks to measure students' performance on the following higher-order skills: Analytic Reasoning and Evaluation, Writing Effectiveness, Writing Mechanics, and Problem Solving. It consists of two task types: Performance Tasks and Analytic Writing (not used in the CWRA).

Performance Tasks ask students to complete a "real-life" activity, such as preparing a memo or policy recommendation. Each Performance Task contains a Document Library—a range of information sources which may include letters, memos, summaries of research reports, newspaper articles, maps, photographs, diagrams, tables, charts, and interview notes or transcripts. These tasks often require students to integrate evidence from different sources; distinguish rational from emotional arguments and fact from opinion; understand data in tables and figures; deal with inadequate, ambiguous, or conflicting information; spot deception and holes in 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. Students are allotted 90 minutes to complete their answers for Performance Task questions.

The Analytic Writing Task contains two types of essay prompts: a *Make-an-Argument* section in which students are asked to take a position and craft a persuasive argument and a *Critique-an-Argument* section which requires students to identify and describe logical flaws in a given argument. Both sections measure a student's skill in articulating complex ideas, examining claims and evidence, supporting ideas with relevant reasons and examples, sustaining a coherent discussion, and using standard written English. Students have 75 minutes to complete the Analytic Writing Task, of which 45 minutes are allotted to the *Make-an-Argument* prompt and 30 minutes are allotted to the *Critique-an-Argument* prompt. The Analytic Writing Task is only used with the CLA and CCLA and is not a part of the CWRA.

Schools typically collect cross-sectional data, with a sample of entering students in the fall and a sample of exiting students in spring of the same academic year. These schools receive two reports of their results. In the fall, schools receive an interim report reflecting data on entering students' scores, both within the school and across institutions administering the CLA. In the spring, schools receive a full report that includes data from both samples of students within the school and across CLA schools. For CWRA schools, the reports include statistics on students' college readiness, comparisons across high schools, and effect sizes, which reflect exiting high school seniors' estimated performance levels relative to entering college freshman. For CCLA schools, the final reports include student growth estimates, effect sizes, and comparisons across community colleges. For CLA schools, final reports include value-added estimates of the institutions' contributions to their students' learning, summary statistics, and comparisons both within and across college.

CLA TASKS

How are CLA tasks developed?

Task development is an iterative process. A team of researchers and writers generate ideas for Analytic Writing prompts and Performance Task storylines, and then contribute to the development and revision of the prompts and Performance Task documents. The researchers who develop CLA tasks have backgrounds in psychometric measurement and writing, as well as extensive experience with test development and writing evaluation.

Analytic Writing Tasks generally take less time to develop than Performance Tasks. For these tasks, multiple prompts are generated, revised and pre-piloted, and those prompts that elicit good critical thinking and writing responses during pre-piloting are further revised and submitted for more extensive piloting.

Performance Task development is a much more involved process, during which CAE takes care to ensure that the task contains sufficient information to permit multiple reasonable solutions. Task developers craft documents to allow for the presentation of information in multiple formats (e.g., tables, figures, news articles). Throughout Performance Task development, CAE establishes and revises a list of the intended content from each document. This list ensures that the documents clearly convey that content and that no additional and unintentional content is imbedded in any of the documents. During revision, task developers adjust the documents' content to guarantee that students could arrive at approximately three or four different conclusions, each informed by the evidence provided in the task documents. Typically, some conclusions are designed to be better supported than others. Questions for the Performance Task are also drafted and revised during document development. The questions are designed such that the first questions prompt the student to read and attend to multiple sources of information in the documents, and later questions require the student to draw conclusions and justify them using evidence from the Document Library.

After several rounds of revision, the most promising tasks are selected for pre-piloting. Task developers examine student responses to identify what pieces of information are unintentionally unclear in the

Performance Task documents or Analytic Writing prompts and what pieces of information are inadvertently included in the documents that should be removed. After revision and additional pre-piloting, the tasks that best elicit the intended types and ranges of student responses are selected for full piloting. During piloting, students complete both an operational task and one of the pilot tasks. At this point, draft scoring procedures are revised and tested in grading the pilot responses, and final revisions are made to the tasks to ensure that the task is eliciting the types of responses intended.

Why are both Performance Tasks and Analytic Writing Tasks necessary?

CLA scores reflect a holistic assessment of the higher order skills of critical thinking, analytic reasoning, written communication, and problem solving. All Performance Tasks and Analytic Writing Tasks require the use of all of these skills, but in different proportions. For example, Analytic Writing Tasks strongly emphasize written communication while Performance Tasks elicit greater use of problem solving skills. Similarly, prompts within each task type vary slightly in which skills they draw upon most.

With the exception of the CWRA, which relies solely on Performance Tasks, students are randomly assigned to a task type and then to a prompt within that task, so each student only answers a small portion of the full complement of CLA prompts. By using this "matrix sampling" strategy, institutions reduce the testing burden on individual students and benefit from the full breadth of the task types.

A document describing the [Architecture of the CLA Tasks] provides additional detail on CLA/CWRA scoring and task development, and is available on the [CLA website].

SCORING

Can you describe the CLA scoring rubrics?

Until the 2010-11 administration, the CLA tasks were scored using varying numbers of holistic and analytic items, with rubrics specifically tailored to each prompt. Because each rubric had a different raw score range, scores were converted to a common scale to allow combining scores from different tasks to compute a school's average score for each task type, as well as a total average score across the task types.

In fall 2010, the CLA introduced a new scoring rubric, which allows schools to not only examine performance across tasks, but also to examine performance by the skills necessary for students to perform well on the assessment. The new scoring rubrics contain subscore categories that reflect Analytic Reasoning and Evaluation, Writing Effectiveness, Writing Mechanics, and Problem Solving (used in Performance Task scoring only).

Evidence of each subscore is elicited in different ways within the context of each task type. For example, in the context of the Performance and Critique-an-Argument Tasks, the Analytic Reasoning and Evaluation category would include interpreting, analyzing, and evaluating the quality of information provided in the Document Library. In the Make-an-Argument Task, Analytic Reasoning and Evaluation involves stating a position, providing valid reasons to support the writer's position, and considering and possibly refuting alternate viewpoints.

Subscores are assigned on a scale of 1 (lowest) to 6 (highest). For all task types, blank or entirely off-topic responses are flagged for removal from results. Because each prompt may have differing possible arguments or relevant information, scorers receive prompt-specific guidance in addition to the scoring rubrics.

You may view the scoring rubrics for each task type online. The CLA and CCLA rubric is available at http://www.collegiatelearningassessment.org/files/CLAScoringCriteria.pdf, and the CWRA rubric is available at http://www.collegiatelearningassessment.org/files/CWRAScoringCriteria.pdf.

Please note that subscores are not adjusted for difficulty like the overall CLA scale scores, and therefore are not directly comparable to each other. The scores are intended to facilitate criterion-referenced interpretations, as defined by the rubrics.

How are CLA tasks scored?

Through the 2007-08 administration, all scoring was conducted by trained graders. Between fall 2008 and spring 2010, a combination of automated and human scoring was used, and since fall 2011 the CLA scoring process has been almost exclusively automated. Automated scoring helps to increase scoring accuracy, reduce the amount of time between a test administration and report delivery, and reduce costs.

The CLA now relies primarily on Pearson's Intelligent Essay Assessor (IEA) for scoring. IEA is the automated scoring engine developed by Pearson Knowledge Technologies to evaluate the meaning of text, not just writing mechanics. Pearson has trained IEA for the CLA using a broad range of real CLA responses and human-generated scores to ensure its consistency with human scorers. Thus, human scorers remain the basis for scoring the CLA tasks.

The automated essay scoring technique that the CLA uses is known as Latent Semantic Analysis (LSA), which extracts the underlying meaning in written text. LSA uses mathematical analysis of about 500 student responses per prompt and the collective expertise of human scorers, and applies what it has learned from the expert scorers to previously unscored student responses.

CAE used an array of Performance and Analytic Writing Tasks to compare the accuracy of human versus automated scoring. For twelve of the thirteen tasks examined, IEA scores agreed more often with the average of multiple experts (r=.84-.93) than two experts agreed with each other (r=.80-.88). These results suggest that computer-assisted scoring is as accurate as—and in some cases, more accurate than—expert human scorers (Elliot, 2011).

Though the majority of scoring is handled by IEA, some responses are scored by trained human scorers. IEA identifies unusual responses, which are automatically sent to the human scoring queue. In addition, ten percent of responses are scored by both IEA and humans in order to continually evaluate the quality of scoring.

For more information about computer-assisted scoring on the CLA, please see *Computer-Assisted Scoring of Performance Tasks for the CLA and CWRA*, available on our website at http://www.collegiatelearningassessment.org/files/ComputerAssistedScoringofCLA.pdf.

How are graders trained and evaluated?

All scorer candidates undergo rigorous training in order to become certified CLA scorers. Training for the Make-an-Argument prompts and Performance Tasks takes place over two days and training for the Critique-an-Argument prompts lasts one day. All training includes an orientation to the prompt and scoring rubrics/guides, repeated practice grading a wide range of student responses, and extensive feedback and discussion after scoring each response.

After participating in training, graders complete a reliability check where they score the same set of student responses. Scorers with low agreement or reliability (determined by comparisons of raw score means, standard deviations and correlations among the scorers) are either further coached or removed from scoring.

SCALING PROCESS

What is the procedure for converting raw scores to scale scores?

For each task, raw subscores are summed to produce a raw total score. Because not all tasks have equal levels of difficulty, raw total scores from the different tasks are converted to a common scale of measurement to reflect comparable levels of proficiency across tasks. Once converted, a given CLA scale score indicates approximately the same percentile rank regardless of the task on which it was earned. This feature of the CLA scale scores allows combining scores from different tasks to compute a school's mean scale score for each task type as well as a total average scale score across types.

To convert the raw scores to scale scores, CAE uses a linear scale transformation. This process results in a scale score distribution with the same mean and standard deviation as the SAT scores of the freshmen who took that measure. This type of scaling preserves the shape of the raw score distribution and maintains students' relative standing on a given task. For example, the student with the highest raw score on the task will also have the highest scale score on that task, the student with the next highest raw score will be assigned the next highest scale score, and so on.

This type of scaling makes it such that a very high score earned on the task (not necessarily the highest possible score) corresponds approximately to the highest SAT (or converted ACT) score of any freshman who took that task. Similarly, a very low score earned on a task would be assigned a scale score value that is close to the lowest SAT (or converted ACT) score of any freshman who took that task. On the rare occasion that a student achieves exceptionally high or low raw scores, this scaling procedure may produce scale scores outside the normal SAT (Math + Verbal) score range of 400 to 1600. Prior to the spring of 2007, scores were capped at 1600 (the maximum allowable on the SAT), though capping was discontinued starting in fall 2007.

Do scaling equations change with each administration?

From fall 2006 to spring 2010, CAE used the same scaling equations for each assessment cycle in order to facilitate year-to-year comparisons.

With the introduction of new scoring criteria in fall 2010, however, raw scores are now on a different scale than in previous years, which made it necessary to revise the scaling equations. Under the new scaling equations, fall 2010 responses tended to receive somewhat lower scores than responses of the same quality would have received in previous years.

If you are interested in drawing comparisons between the average CLA scale scores in your current institutional report and those reported prior to fall 2010, we encourage you to use the appropriate equation below to convert pre-fall 2010 scale scores to current scale scores. The correlation between institution average scores on the old and new score scales is 0.99, and this equation characterizes the strong linear relationship between those scores. The equation can apply to all institution-level score types: Total, Performance Task, Analytic Writing Task, Make-an-Argument, and Critique-an-Argument.

CLA & CCLA: $score_{new} = 102.29 + (0.8494 \cdot score_{old})$

CWRA: $score_{new} = 98.08 + (0.8704 \cdot score_{old})$

VALUE-ADDED SCORING

What do my school's value-added scores on the CLA mean?

When the average performance of seniors at a school is substantially better than expected, this school is said to have high "value added." To illustrate, consider several schools admitting students with similar average performance on general academic ability tests (e.g., the SAT or ACT) and on tests of higher-order skills (e.g., the CLA). If, after four years of college education, the seniors at one school perform better on the CLA than is typical for schools admitting similar students, one can infer that greater gains in critical thinking and writing skills occurred at the highest performing school.

Note that a low (negative) value-added score does not necessarily indicate that no gain occurred between freshman and senior year; however, it does suggest that the gain was lower than would typically be observed at schools testing students of similar entering academic ability.

Value-added scores are placed on a normalized (z-score) scale and assigned performance levels. Schools that fall between -1.00 and +1.00 are classified as "near expected," between +1.00 and +2.00 are "above expected," between -1.00 and -2.00 are "below expected," above +2.00 are "well above expected," and below -2.00 are "well below expected." Value-added estimates are also accompanied by confidence intervals, which provide information on the precision of the estimates; narrow confidence intervals indicate that the estimate is more precise, while wider intervals indicate less precision.

Please note that only four-year colleges participating in the both windows of the CLA receive value-added scores. CCLA and CWRA schools instead receive effect sizes, which reflect the standardized differences in scores between entering and exiting students, using a school's standard deviation for entering students. An effect size of 0 indicates no difference between entering and exiting students. Positive effect sizes indicate that scores of exiting students are higher than those of entering students, with larger effect sizes corresponding to larger score differences. Effect sizes of greater than 0.50 are generally considered large.

CWRA schools also receive deviation scores to estimate college readiness by comparing a high school's exiting students to college freshman. Deviation scores show the differences between observed and expected scores and are reported in standard deviation units. These scores are placed on a standard-normal (z-score) scale as with the CLA value-added scale above.

What value-added model does the CLA use?

Through spring 2009, the CLA estimated value added as the difference between freshman and senior deviation scores through an ordinary least squares (OLS) regression model. Beginning in fall 2009, the CLA moved to an enhanced regression model known as hierarchical linear modeling (HLM), which accounts for CLA score variation within and between schools. Under the new model, a school's value-added score indicates the degree to which the observed senior average CLA score meets, exceeds, or falls below expectations established by the senior average Entering Academic Ability (EAA) score and the average CLA performance of freshmen at that school, which serves as a control for selection effects not covered by EAA. Only students with EAA scores—SAT Math + Verbal, ACT Composite, or Scholastic Level Exam (SLE) scores converted to the SAT scale—are included in institutional analyses.

The decision to move from an OLS to HLM model was made after analyses showed that the two methods produce similar results. Correlations between the value-added scores resulting from the two approaches were .79 in the 2006-07 administration and .72 in the 2007-08 administration. Reliability estimates, however, were higher for the newer model than the original. Average split-sample reliabilities were .81 (HLM) and .73 (OLS) for 2006-07, and .75 (HLM) and .64 (OLS) in 2007-08. Year-to-year value-added score correlations also increased with the new approach (.58) from the original (.32). The HLM model, therefore, is more efficient because, when the number of tested students is held constant, scores from the new approach are more precise within a given year and are also more realistically stable across years. The HLM model also provides school-specific indicators of value-added score precision, which improve the interpretability of scores.

For more information about the difference between the OLS and HLM models, as well as the rationale for moving to the newer model, please see Improving the Reliability and Interpretability of Value-Added Scores for Post-Secondary Institutional Assessment Programs (Steedle, 2010a).

How can I calculate my value-added score?

Institutions may want to conduct their own analyses in which, for example, they calculate the value-added scores within certain subpopulations of students for whom they have conducted in-depth sampling. To calculate these scores, you need:

- Samples of entering and exiting students with both CLA and EAA scores. This information is available in the Student Data File, which is distributed to institutions with each administration's results.
- The estimated parameters for the value-added model, which are provided in the appendices to your institutional report. The parameters for the 2011-12 administration are in the table below.

	Intercept	Senior EAA Slope	Freshman CLA Slope	Value-Added Score Standard Deviation
Total Score	341.48	0.46	0.31	50.11
Performance Task	331.73	0.53	0.25	60.22
Analytic Writing Task	372.61	0.38	0.36	50.48
Make-an-Argument	350.18	0.35	0.40	52.82
Critique-an-Argument	390.98	0.46	0.27	58.51

The steps for calculating your value-added score are listed below. Please note that, while the equations in this process are for finding the Total CLA Score value-added, the same equation can be used for individual task types, as well. Simply insert the "Total Score" parameters with the individual task types in the table above.

1. Refer to your CLA Student Data File to identify your subgroup sample of interest. The sample must contain freshmen and seniors with CLA scores (Performance Task or Analytic Writing Task) ***and*** EAA (entering academic ability).
2. Using the Student Data File, compute:
   a. The mean EAA score of seniors (exiting students) in the sample
   b. The mean CLA score of freshmen (entering students) in the sample
   c. The mean CLA score of seniors (exiting students) in the sample
3. Calculate the senior subgroup sample's expected mean total CLA score using the equation below:
   a. The expected senior mean CLA total score = (*Total Score Intercept* from parameters table) + (*Total Score Senior EAA Slope* from parameters table)*(senior mean EAA) + (*Total Score Freshman CLA Slope* from parameters table)*(freshman mean CLA)
4. Use your expected score to calculate your subgroup sample's value-added score in standard deviation units, using the following equation:
   a. Value-added score = [(observed senior mean CLA total score) – (expected senior mean CLA total score)] / (*Total Score Value-Added Score Standard Deviation*)

ANALYSIS

What is the process for averaging students' scores for comparison and reporting?

To be included in the calculation of a mean score for a school, students must:

- be in the correct class year (verified by the registrar);
- have either an ACT Composite, or SAT Math and SAT Verbal/Critical Reading score, or SLE score;
- and have a complete CLA task score (either Performance Task or Analytic Writing Task).

The total scale score is the mean of the Performance Task and Analytic Writing Task scale scores for those students with ACT/SAT/SLE scores and who are in the correct class year.

Does CLA analysis account for ceiling effects?

No school's averages approach the theoretical maximum of scaled CLA scores. There are, however, individual students who have, in the past, achieved a maximum scale score on the CLA, as a function of exceptional performance. Historically, we capped the distribution at 1600 (the maximum of the SAT Verbal/Critical Reading + SAT Math). This did impact the scores of a very small percentage of students. After researching this further, we opted to lift the cap, starting in fall 2007.

Does the CLA correct for range restriction?

Correcting for range restriction is not necessary here because the institution is the unit of analysis, and we don't have a range-restricted population of institutions. Summary statistics of SAT scores for students sampled in the CLA are similar to national figures. Specifically, looking at the 2008 estimated median SAT (or ACT equivalent) of the freshman class across 1,398 four-year institutions in the U.S, we find a minimum of 726, mean of 1057, maximum of 1525, and standard deviation of 127. Across CLA schools (spring 2011, n=190) for the same variable we find a minimum of 726, mean of 1041, maximum of 1345, and standard deviation of 106 (College Results Online, 2010).

CONTROLLING FOR ENTERING ACADEMIC ABILITY (EAA)

How does the CLA "crosswalk" between the ACT and the SAT?

If a participating institution collects ACT scores instead of SAT scores, they are converted to the SAT's scale of measurement using a standard crosswalk. The maximum ACT score of 36 corresponds to the SAT (Math + Verbal/Critical Reading) maximum of 1600, an ACT score of 35 corresponds to 1560, and so forth (ACT, 2008).

A full crosswalk is printed in the sample institutional report available on our website, and is also provided to participating institutions in both the fall and spring reports. The correlation between ACT Composite and SAT Math + SAT Verbal/Critical Reading has been shown to be as high as .92 (Dorans & Schneider, 1999).

How strong is the correlation between the SAT/ACT and the CLA?

At the student level of analysis, correlations between SAT/ACT and CLA scores are .44 and .46 for freshmen and seniors, respectively on the Analytic Writing Tasks (Klein, Benjamin, Shavelson, & Bolus, 2007). Student-level correlations on Performance Tasks are .56 and .54.

At the institutional level of analysis, correlations for freshmen and seniors on the Analytic Writing Tasks are .79 and .83, respectively. Institution-level correlations on Performance Tasks are .97 and .88.

If the SAT/ACT and CLA are so closely correlated, why can't the SAT/ACT be used as a substitute for freshmen scores? That is, why test freshmen at all?

The high correlation between the tests does not imply that they measure the same thing or are designed with the same purpose in mind. For instance, the CLA, unlike the SAT and ACT, is a test that is meant to be taught to, and the principal goal of the CLA is to facilitate teaching and learning (Benjamin, Chun, & Jackson, 2009).

SAT/ACT scores simply allow us to control for general cognitive ability. SAT/ACT scores, if used as a substitute for freshmen CLA scores, would not account for first-year students' entering level analytic reasoning, critical thinking, problem solving, and written communication skills (as measured by the CLA).

What evidence do you have that SLE scores are equivalent to SAT scores to control for entering academic ability?

The Scholastic Level Exam (SLE) is a short-form cognitive aptitude test produced by Wonderlic (a commercial test provider). The CLA uses SLE scores to define a student's Entering Academic Ability (EAA) when SAT or ACT scores are not available or applicable (e.g., for entering high school students).

In spring 2006, over 1,150 students (seniors at four-year colleges and universities and exiting students at community colleges) took the SLE in addition to either a 90-minute Performance Task or a 75-minute Analytic Writing Task as part of the CLA. Registrar offices supplied ACT and/or SAT scores for these students. Students were given 12 minutes to complete the 50-item SLE online. The mean total score (sum of all items, worth one point apiece) was 38, with a standard deviation of 8.

Student-level correlations between the SLE total score and ACT Composite, SAT Verbal, SAT Math, and SAT Composite Equivalent scores were as follows:

.68	SLE and ACT Composite
.68	SLE and SAT Verbal/Critical Reading
.66	SLE and SAT Math
.70	SLE and SAT Composite Equivalent

Across 24 schools where at least 10 students had both SLE and SAT or ACT scores, the school-level correlation between the mean SLE total score and mean SAT Composite Equivalent score was .92.

Wonderlic reports that SLE scores, as a measure of cognitive ability, are stable over time. SLE scores need not be age-adjusted for students between ages 15 and 29.

CORRELATIONS WITH OTHER MEASURES

To what degree is the National Survey of Student Engagement (NSSE) correlated with the CLA?
Correlations between the National Survey of Student Engagement (NSSE) and CLA were explored using data from the CLA feasibility study. Findings were presented at the 2004 annual meeting of the American Educational Research Association, and published in *Research in Higher Education* (Carini, Kuh, & Klein, 2006). The researchers found statistically significant—but small—correlations between CLA outcomes and student engagement scores. Partial correlations between CLA outcomes and student engagement scales were .10 or higher for level of academic challenge, supportive campus climate, reading and writing, quality of relationships, institutional emphases on good practices, and self-reported general education gains. None of the CLA-engagement partial correlations was negative, and they were also slightly higher than GRE-engagement correlations. An abstract of this article follows:

> This study examines (1) the extent to which student engagement is associated with experimental and traditional measures of academic performance, (2) whether the relationships between engagement and academic performance are conditional, and (3) whether institutions differ in terms of their ability to convert student engagement into academic performance. The sample consisted of 1058 students at 14 four-year colleges and universities that completed several instruments during 2002. Many measures of student engagement were linked positively with such desirable learning outcomes as critical thinking and grades, although most of the relationships were weak in strength. The results suggest that the lowest-ability students benefit more from engagement than classmates, first-year students and seniors convert different forms of engagement into academic achievement, and certain institutions more effectively convert student engagement into higher performance on critical thinking tests.

Are there linkages or relationships between your test and any standardized placement test (e.g., a test used to determine what initial math or English course a freshman should take) such that the placement test could serve as a control for the entering ability of students?
To date, we have not conducted research to determine whether any linkages or agreements between the CLA and various standardized placement tests that would determine an initial freshman course exist. That being said, some participating institutions are utilizing the CLA in a pre/post fashion to determine the efficacy of certain programs or courses for entering students.

RELIABILITY

What is the reliability of the CLA?

The reliability of CLA scores is assessed from multiple perspectives during each administration. Since fall 2011, all prompts have been graded through an automatic scoring process, though approximately 10 percent of the responses are also scored by trained graders to continually evaluate the quality of scoring and confirm consistency across scorers. For details about the consistency of scorers, please see *How are Tasks Scored* (p. 4).

VALIDITY

Do you have any evidence of construct validity?

In the fall semester of 2008, CAE (CLA) collaborated in a construct validity study with ACT (CAAP) and ETS (MAPP) to investigate the construct validity of these three assessments (Klein et al., 2009). Construct validity refers to whether an assessment measures the particular skill (i.e., construct) that it purports to measure and 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, Liu, et al., 2009) can be found on CAE's website at http://www.cae.org/content/pdf/TVS_Report.pdf.

What about the face validity of your measure?

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, CLA tasks must emulate the critical thinking and writing challenges that students will face outside the classroom. These characteristics of the CLA were vetted by a sample of 41 college professors selected to be representative of faculty from a wide range of institutions (Hardison & Vilamovska, 2008). After reviewing CLA Performance Tasks in depth and reading a range of student responses, these professors completed a questionnaire to express their perceptions of the tasks.

As shown in Figure 1, results indicate that the professors perceived the Performance Tasks to be good assessments of critical thinking, writing, problem solving, and decision making. Responding on a 1-5 scale, professors felt, for example, that the CLA measures what it intends to measure (Mean 4.14, SD 0.46); it measures important skills that college graduates should possess (Mean 4.70, SD 0.53); students need good critical thinking skills to do well on the task (Mean 4.60, SD 0.46); and students who do well on the task would also perform well in a job requiring good written communication (Mean 4.20, SD 0.83) or decision-making (Mean 4.10, SD 0.70). Respondents also agreed, after viewing the tasks, that college seniors should perform better on this task than college freshman (Mean 4.70, SD 0.48).

We also encourage you to view the CLA yourself. To access a retired CLA Performance Task, please visit http://starttest.com/7.0.0.1/programs/clacross/Practice%20Test%20Page.htm. You will see the same testing interface that students use. The demonstration will allow you to view one Performance Task, though students typically take one of approximately eight different prompts from the available Performance Tasks.

Sample Analytic Writing Tasks (Make-an-Argument and Critique-an-Argument) are also available on our website at: http://www.cae.org/content/pro_collegiate_sample_measures.htm.

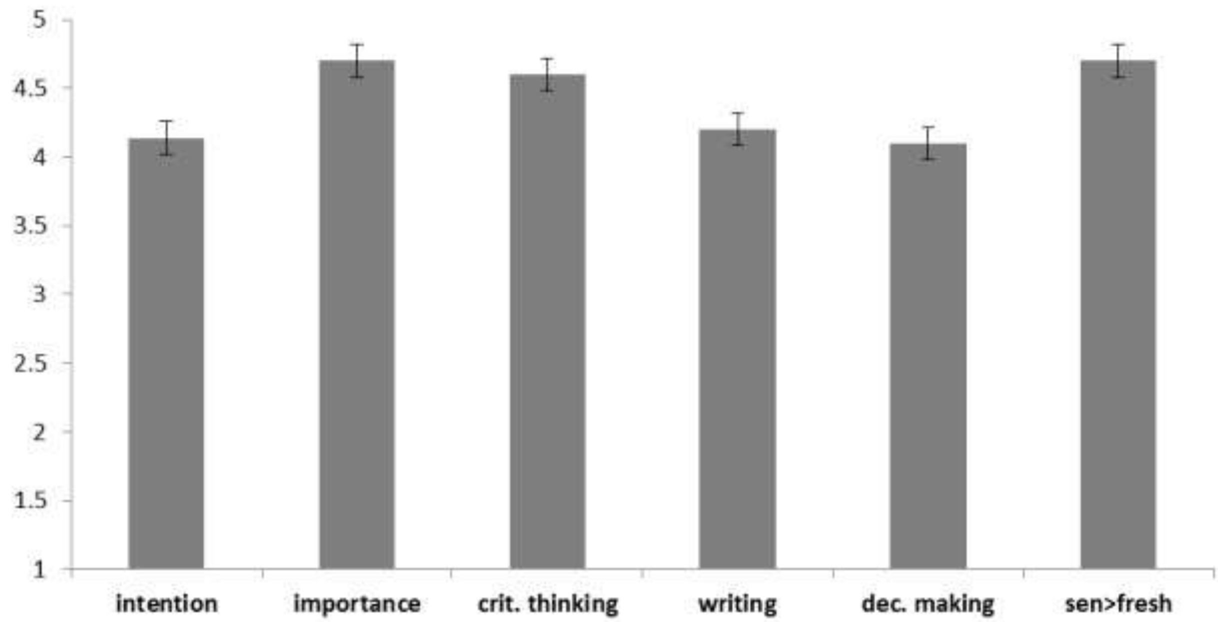

Figure 1: Average face validity assessments of the CLA

STUDENT EFFORT

We are concerned that students won't devote sufficient effort to the CLA and that our CLA institutional results will suffer as a result. Do you control for student effort?

The CLA does not control for self-reported student effort, but has conducted some research on the role that motivation plays in CLA achievement. Analyses of the relationship between Performance Task scores and self-reported effort suggest that, controlling for entering academic ability, student effort only explains about three to seven percent of the variance in school-level scores (Klein, et al., 2007).

Additional research, presented at the 2010 Annual Meeting of the American Educational Research Association, focused on the relationship between incentives, motivation, and CLA performance. Using the Student Opinion Survey (SOS)—a motivation scale that measures a student's effort and belief that performing well is important—CAE found that (after controlling for average entering academic ability) motivation was a significant predictor of CLA scores on the student level, but not on the school level (Steedle, 2010b).

OTHER FACTORS OF INTEREST

Are there differences in scores by sex? By racial/ethnic group? By school characteristics?
Whether test-takers in one demographic subgroup (e.g., female or African American students) perform as well as other test-takers with similar ability levels in another subgroup (e.g., male or Hispanic/Latino students) is an important question because the presence of such differences might suggest the presence of subgroup bias. To answer this question, CLA researchers conducted a series of regression analyses on over 17,000 first-year students and over 10,900 seniors taking the CLA in fall 2005 and spring 2006 to determine whether student ability (as measured by SAT or ACT scores), race/ethnicity, gender or primary language spoken (other than English) could explain more variation in student scores than student ability alone.

Results from a study (Steedle, 2011) show that after controlling for EAA and average freshman CLA scores, at the institutional level, many variables (e.g., public/private, selectivity rate, size, retention rate, etc..) were not significant predictors of average senior CLA scores at the institution level.

Is there an interaction between performance task "topic" and a student's major?
We have looked for, but have not found, any interaction between task topic and student major. Analyses using data from college seniors who took completed a Performance Task in spring 2007 show that, controlling for a combination of task topic, student academic major, SAT scores, and the interaction of task and major, only the SAT variable contributed to predictive accuracy or value added (Klein, Freedman, Shavelson, & Bolus, 2008). CAE is currently investigating this relationship again with more recent data, and will release the results from the new analysis in the near future.

What is the relationship between CLA scores and time spent on CLA tasks?
There is a moderate positive correlation between CLA scores and time spent on CLA tasks and between SAT scores and time spent on CLA tasks. Most students need up to 90 minutes to fully address the Performance Task and up to 75 minutes to fully address the Analytic Writing Task. This relationship not only normal, it is to be expected. Good responses tend to be longer and therefore take longer to compose. It does not, however, mean that students cannot achieve a high score on the CLA with a relatively brief response.

Why does CAE recommend a sample size of 100 students per cohort?
CAE recommends sampling 100 (or more) students in each cohort an institution tests in order to increase the precision of value-added estimates. Along-side value-added scores, CAE provides confidence intervals for CLA schools that have tested both entering and exiting students in a given academic year. Analysis of the precision of the current CLA value-added model, using data from 2007-2008, shows that schools testing larger numbers of students obtain more precise value-added estimates (Steedle, 2010a). Figure 2 below shows that the size of the 95% confidence interval decreases sharply as sample size increases toward 100 students

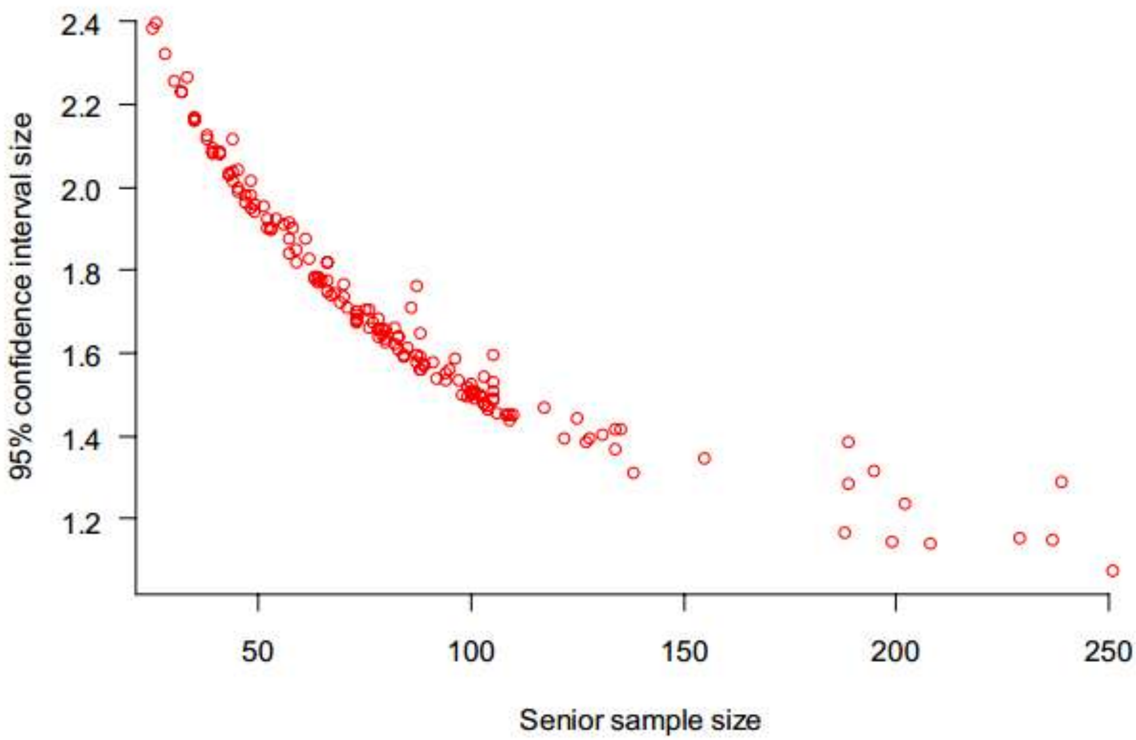

Figure 2: Relationship between 95% confidence interval size and senior sample size for the 2007-2008 data.

ADDITIONAL TECHNICAL INFORMATION ABOUT THE CLA

A list of documents, with links to their locations online, is available below.

Background and general information about the CLA:

- An Approach to Measuring Cognitive Outcomes Among Higher Education Institutions (2005)
- The Collegiate Learning Assessment: Facts and Fantasies (Klein, et al., 2007)
- The Collegiate Learning Assessment's Place in the New Assessment and Accountability Space (Benjamin, Chun, & Jackson, 2009)

- Returning to Learning in an Age of Assessment: Introducing the Rationale of the Collegiate Learning Assessment (Benjamin et al., 2009)
- General information from the CLA, CCLA, and CWRA sections of our website

CLA tasks and scoring:

- Architecture of the CLA Tasks
- CLA/CCLA and CWRA Scoring Rubrics
- Computer-Assisted Scoring of Performance Tasks for the CLA (Elliot, 2011)

Factors influencing CLA outcomes and other variable interactions:

- Incentives, Motivation, and Performance On a Low-Stakes Test of College Learning (Steedle, 2010b)
- Assessing School Effectiveness (Klein, et al., 2008)

Incorporating teaching and learning into the CLA:

- CLA in the Classroom, the CLA program that focuses on enhancing teaching and learning through Performance Task Academies and other education resources
- Teaching to a Test Worth Teaching To (Hersh, 2008)

Methods and structural approaches to the CLA:

- Improving the Reliability and Interpretability of Value-Added Scores for Post-Secondary Institutional Assessment Programs (Steedle, 2010a)
- The Lumina Longitudinal Study: Summary of Procedures and Findings (Klein, Steedle, & Kugelmass, 2009) a comparison of the longitudinal and cross-sectional approach to the CLA

Validity of the CLA:

- The Collegiate Learning Assessment: Setting Standards for Performance at a College or University (Hardison & Vilamovska, 2008)
- Test Validity Study Report (Klein, Liu, et al., 2009)

Other research and analyses using the CLA:

- Academically Adrift: Limited Learning on College Campuses (Arum & Roksa, 2011)
- Learning to Reason and Communicate in College: Initial Report of Findings from the CLA Longitudinal Study (Arum, Roksa, & Velez, 2008)

All articles and resources listed above, and more, are located on the main page of the CLA website at http://www.collegiatelearningassessment.org/ or in the Articles and Research Section of the CLA website at http://www.cae.org/content/pro_collegiate_reports_publications.htm.

REFERENCES

ACT. (2008). ACT-SAT concordance. Retrieved July 21, 2011, 2011, from http://www.act.org/aap/concordance/

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Arum, R., Roksa, J., & Velez, M. (2008). *Learning to reason and communicate in college: Initial report of findings from the CLA longitudinal study*. Brooklyn, NY: Social Science Research Council.

Benjamin, R., Chun, M., Hardison, C. M., Hong, E., Jackson, C., Kugelmass, H., et al. (2009). *Returning to learning in an age of assessment: Introducing the rationale of the Collegiate Learning Assessment*. New York: Council for Aid to Education.

Benjamin, R., Chun, M., & Jackson, C. (2009). *The collegiate learning assessment's place in the new assessment and accountability space*. New York: Council for Aid to Education.

Campbell, D. T. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education, 47*(1), 1-32.

Dorans, N. J., & Schneider, D. (1999). *Concordance between SAT i and ACT scores for individual students* (No. RN-07).

Elliot, S. (2011). *Computer-assisted scoring for performance tasks for the CLA and CWRA.* New York: Council for Aid to Education.

Hardison, C. M., & Vilamovska, A.-M. (2008). *The Collegiate Learning Assessment: Setting standards for performance at a college or university* (No. PM-2487-1-CAE). Santa Monica, CA: RAND.

Hersh, R. H. (2008). Teaching to a test worth teachng to in college and high school. Council for Aid to Education.

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

Klein, S., Freedman, D., Shavelson, R., & Bolus, R. (2008). Assessing school effectiveness. *Evaluation Review, 32*(6), 511-525.

Klein, S., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., et al. (2009). Test validity study (tvs) report. Supported by the fund for the improvement of postsecondary education. from http://www.voluntarysystem.org/docs/reports/TVSReport_Final.pdf

Klein, S., Steedle, J., & Kugelmass, H. (2009). *CLA Lumina longitudinal study summary findings*. New York: Council for Aid to Education.

Steedle, J. T. (2010a). *Improving the reliability and interpretability of value-added scores for post-secondary institutional assessment programs.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Steedle, J. T. (2010b). *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.

Steedle, J. T. (2011). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education*, 1-16.