cla+

TECHNICAL FAQs

cae | 215 Lexington Avenue, Floor 16 | New York, NY 10016 | cla@cae.org | (212) 217-0700 | @cla_beat

WHAT IS THE DIFFERENCE BETWEEN THE CLA AND CLA+?

When first launched more than a decade ago, the Collegiate Learning Assessment (CLA) pioneered a constructed-response approach to the assessment of higher-order skills. Initially, the CLA was designed to measure an institution's contribution, or value added, to the development of these higher-order skills to its student body, and therefore the institution—not the student—was the primary unit of analysis.

The CLA employed a matrix sampling approach, under which students were randomly distributed either a Performance Task (PT) or an Analytic Writing Task, for which students were allotted 90 minutes and 75 minutes, respectively. The CLA Performance Tasks presented real-world situations in which an issue, problem, or conflict was identified, and students were asked to assume a relevant role to address the issue, suggest a solution, or recommend a course of action based on the information provided in a document library. Analytic Writing Tasks consisted of two components— one in which students were presented with a statement around which they must construct an argument (Make an Argument), and another in which students were given a logically flawed argument that they must then critique (Critique an Argument).

In its original form, the utility of the CLA was limited. Because the assessment consisted of just one or two responses from each student, reliable results were only available at the institutional level, and students' results were not directly comparable. Likewise, reporting for the CLA was restricted to the purposes of its value-added measure, and institutions were not eligible for summary results unless they had tested specified class levels in the appropriate testing windows.

Now, however, CLA+ has both greater utility and more flexibility. CLA+ comprises both a Performance Task and a Selected-Response Question (SRQ) section, with each student receiving both components of the assessment. The inclusion of different item formats not only allows for the assessment of a broad range of content and skills but also affords the opportunity to obtain reliable results at both the institution and student levels. As in its former incarnation, CLA+ allows an institution to compare its student learning results with the learning results at similarly selective institutions, and to use that information to improve teaching and learning. Additionally, CLA+ results at the student level allow for a comparison of how well each individual performed relative to his or her peers at the same class level both within an institution and across the other CLA+ institutions. Unlike the CLA reports, CLA+ reports include data on each of the four standard college class levels, allowing schools to receive results for any class level they test, regardless of the window in which the assessment occurs.

The SRQ section was developed with the intent to assess higher-order cognitive skills rather than the recall of factual knowledge. Similar to the PT, students are presented with a set of questions as well as one or two documents to refer to when answering each question. The supporting documents include a range of information sources, such as letters, memos, photographs, charts, and/or newspaper articles.

In addition to the subscores previously reported for the PT (Analysis and Problem Solving, Writing Effectiveness, and Writing Mechanics), CLA+ uses SRQs to measure students' performance on the following higher-order skills: Scientific and Quantitative Reasoning, Critical Reading and

Evaluation, and Critique an Argument.

As was the case with the CLA, CLA+ requires 90 minutes of testing time, during which students have an hour to complete the PT and half an hour to respond to the 25 SRQs.

Additionally, CLA+ also introduces a new metric in the form of mastery levels. The mastery levels are qualitative categorizations of total CLA+ scores, with cut scores that were derived from a standard-setting study in the fall of 2013. CAE developed a profile for each mastery level, and this information is available in the appendices of each institutional report. The mastery level categories are: Below Basic, Basic, Proficient, Accomplished, and Advanced.

CLA+ TASKS

How are CLA+ tasks developed?

CAE item developers follow a rigorous and structured item development plan when creating new PTs and SRQs. The primary goal is to develop assessment items that are authentic and engaging for the students. This is accomplished through a series of checklists, including whether the students can reasonably craft an argument using only the information that is provided and whether there is enough information to support or refute from multiple perspectives. One of the unique features of CLA+ is that no prior knowledge of any specific content area is necessary in order to perform well on the assessment. Students are assessed on their critical-thinking and written-communication skills, not on how much knowledge they have in subjects such as U.S. history or chemistry.

The documents for both the PTs and the SRQs are presented in the most appropriate format for the scenario. This can include, but is not limited to, an abstract from a journal, tables, charts, graphs, memos, blog postings, newspaper articles, maps, and reports. Throughout development, CAE item developers outline, write, and revise the content from each document within a PT or SRQ section. This process ensures that the documents cover all of the necessary information and that no additional or unintentional content is imbedded in or missing from the documents. CAE editors review initial drafts of the tasks and provide feedback to the developer for revisions.

For the PTs specifically, item developers are instructed to create scenarios where there is more than one possible conclusion, solution, or recommendation. Each possible outcome is supported by at least some evidence provided in the documents. Typically, some of the possible conclusions are designed to be better supported than others. However, there is always enough material in the document library to fully support any position allowed by the scenario. As long as the student's response aligns to the criteria in the appropriate range of the scoring rubric, that student can still earn the highest scores.

The SRQ section, like the PT, represents a real-world scenario and problem. Students are expected to answer questions that require them to critically read and evaluate a passage or situation, use scientific and/or quantitative reasoning, and identify logical fallacies in an argument. These types of questions, therefore, require students to think at a deeper level than the traditional recall-and-recognition questions that are seen on many standard multiple-choice assessments.

After several rounds of revision between the developer and one or more of CAE's editors, the most promising tasks are selected for pilot testing. In each testing window, there is one PT that is in a pilot phase. Once enough responses to the pilot PT are collected, scaling and equating equations can be created to make scores on the pilot PT equivalent to scores on each of the other PTs. Additionally, 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. More details on the scaling and equating methods are presented below.

For the SRQ section, a classical item analysis is conducted after the pilot testing to determine whether further revisions are necessary before the item becomes operational. Items are examined in terms of item discrimination and item difficulty. A point-biserial correlation is computed to determine the relationship between the item score (correct versus incorrect) and the total test score. This value is often referred to as the item discrimination index. A high correlation between the item score and the total test score is an indication that the item does well at discriminating between students with low test scores and students with high test scores, and that the item is therefore appropriate for the test. The item difficulty, called a "p-value," is the proportion of students that answered the item correctly. The p-value is examined to ensure that there is sufficient range in terms of item difficulty, meaning there should not be too many items that are either very difficult or very easy. The items that are too difficult or too easy tend to not have satisfactory point-biserial correlations because too many responses are correct (or incorrect) and a statistical relationship can therefore not be established. Operational items in the CLA+ bank have p-values between .30 and .80 and a point-biserial correlation of at least .10.

The item developers, editors, and measurement scientists who develop CLA+ tasks have varied backgrounds including history, English, mathematics, psychology, and psychometrics. Over the years of developing the CLA and CLA+, the team now has extensive experience with test development and writing evaluation.

What is the benefit of including different item formats (Performance Tasks and Selected-Response Questions) in the assessment?

As mentioned in the introduction, prior to CLA+, the assessment was only valid and reliable at the institutional level. CLA clients often asked if the student results could be used to make decisions about performance at the individual student level. CAE recommended against using the scores as individual assessments because the reliability at the individual level was not established.

In order to increase the reliability of the CLA scores for individual students, more data needed to be collected from each student. There were several different models that could have been employed, including administering more than one PT to each student. However, due to the desire to limit the amount of time students spend testing, CAE decided to develop CLA+ with the traditional performance-based PT as the anchor and a set of 25 SRQs, which assess the same construct as the PT (analytic reasoning and problem-solving). These SRQs boost the reliability at the individual student level significantly while keeping the total testing time the same as the original CLA.

Additionally, both PTs and SRQs are capable of measuring critical-thinking skills. Each section has strengths and weaknesses and the arrangement of the two different format types creates a balance of strengths relative to weaknesses in terms of content coverage, reliability and validity evidence, and scoring objectivity and efficiency.

SCORING

Can you describe the CLA+ scoring rubrics?

The CLA+ scoring rubric for the PTs consists of three subscores: Analysis and Problem Solving (APS), Writing Effectiveness (WE), and Writing Mechanics (WM). Each of these subscores is scored from a range of 1-6, where 1 is the lowest level of performance and 6 is the highest level of performance, with each score pertaining to specific response attributes. For all task types, blank or entirely off-topic responses are flagged for removal from the results and do not receive a score.

APS measures a student's ability to make a logical decision or conclusion (or take a position) and support it with accurate and relevant information (facts, ideas, computed values, or salient features) from the Document Library.

Writing Effectiveness assesses a student's ability to construct and organize logically cohesive arguments. This is accomplished by strengthening the writer's position by elaborating on facts or ideas (e.g., explaining how evidence bears on the problem, providing examples, and emphasizing especially convincing evidence).

WM evaluates a student's facility with the conventions of standard written English (agreement, tense, capitalization, punctuation, and spelling) and control of the English language, including syntax (sentence structure) and vocabulary.

The CLA+ rubric is available on our website at http://cae.org/claptrubric.

The SRQ section of CLA+ consists of three subsections, each of which has a corresponding subscore category: Scientific and Quantitative Reasoning, Critical Reading and Evaluation, and Critique an Argument. Subscores in these sections are scored according to the number of questions correctly answered, with scores adjusted for the difficulty of the particular question set received. Scores for scientific and quantitative reasoning and critical reading and evaluation can range from 0 to 10, and scores for critique an argument range from 0 to 5.

How does CLA+ scoring work?

All scorer candidates undergo rigorous training in order to become certified CLA+ scorers. Scorer training consists of two to three separate sessions and takes place over several days. A lead scorer is identified for each PT and is trained in person by CAE measurement scientists and editors. Following this training, the lead scorer conducts an in-person or virtual (but synchronous) training session for the scorers assigned to his or her particular PT. A CAE measurement scientist or editor attends this training as an observer and mentor. After this training session, homework assignments are given to the scorers in order to calibrate the entire scoring team. 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. Because each prompt may have differing possible arguments or relevant information, scorers receive prompt-specific guidance in addition to the scoring rubrics. CAE provides a scoring homework assignment for any PT that will be operational before the onset of each testing window to ensure that the scorers are properly calibrated. For pilot PTs, a separate training is first held to orient a lead scorer to the new PT, and then a general scorer training is held to introduce the new PT to the scorers. After participating in training, scorers 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.

During piloting of any new PTs, all responses are double-scored by human scorers. These double-scored responses are then used for future scorer trainings, as well as to train a machine-scoring engine for all future operational test administrations of the PT.

CAE uses Intelligent Essay Assessor (IEA) for its machine scoring. IEA is the automated scoring engine developed by Pearson Knowledge Technologies to evaluate the meaning of a text, not just writing mechanics. Pearson designed IEA for CLA+ using a broad range of real CLA+ responses and scores to ensure its consistency with scores generated by human scorers. Thus, human scorers remain the basis for scoring the CLA+ tasks. However, automated scoring helps increase scoring accuracy, reduce the amount of time between a test administration and reports delivery, and lower costs. The automated essay scoring technique that CLA+ uses is known as Latent Semantic Analysis (LSA), which extracts the underlying meaning in written text. LSA uses mathematical analysis of at least 800 student responses per PT and the collective expertise of human scorers (each of these responses must be accompanied by two sets of scores from trained human scorers), and applies what it has learned from the expert scorers to new, unscored student responses.

Once tasks are fully operational, CLA+ uses a combination of automated and human scoring for its Performance Tasks. In almost all cases, IEA provides one set of scores and a human provides the second set. However, IEA occasionally identifies unusual responses. When this happens, the flagged response is automatically sent to the human scoring queue to be scored by a second human instead of by IEA. For any given response, the final PT subscores are simply the averages of the two sets of scores, whether one human set and one machine set or two human sets.

To ensure continuous human scorer calibration, CAE developed the E-Verification system for the online scoring interface. The E-Verification system was developed to improve and streamline scoring. Calibration of scorers through the E-Verification system requires scorers to score previously-scored results, or "verification papers," when they first start scoring, as well as throughout the scoring window. The system will periodically present verification papers to scorers in lieu of student responses, though they are not flagged to the scorers as such. The system does not indicate when a scorer has successfully scored a verification paper; however, if the scorer fails to accurately score a series of verification papers, he or she will be removed from scoring and must participate in a remediation process. At this point, scorers are either further coached or removed from scoring.

Using data from the CLA, CAE used an array of Performance and Analytic Writing Tasks to compare the accuracy of human versus automated scoring. For 12 of the 13 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).

SCALING PROCESS

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

For the PT, raw subscores are summed to produce a single raw PT total score. The raw PT total score then undergoes a linear transformation to equate it to the scores obtained by our norm population on the original set of PTs. This ensures that PT scores can be compared with each other regardless of which PT was administered or in which year the test was taken.

For the SRQs, the raw subscores first undergo a scaling process to correct for different levels of difficulty of the SRQ sections. A single raw SRQ total score is then computed by taking a weighted average of the SRQ subscores, with weights corresponding to the numbers of items in each of the three SRQ sections. The raw SRQ total score then undergoes a linear transformation to equate it to the scores obtained by our norm population on the original set of SRQs. As with the PTs, this process ensures that SRQ scores can be compared with each other regardless of which SRQ set was administered or in which year the test was taken.

The scaled PT total score and the scaled SRQ total score are then averaged together to create a raw CLA+ total score. The raw total scores undergo a final linear transformation to become scaled CLA+ total scores, again allowing for comparison across all testing windows.

Do scaling equations change with each administration?

Periodically, CAE will update equating equations to ensure continuous comparability across testing windows, and to ensure that PTs are interchangeable and that SRQ sets are interchangeable. Additionally, whenever the norm sample is updated, the equating equations will be updated as well. The next scheduled update to the equating equations is the summer of 2017. However, the norm sample will not be updated at that time.

VALUE-ADDED SCORING

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

CLA+ includes two forms of growth estimates—value-added scores and effect sizes. Value-added scores compare growth on CLA+ within a school to the growth seen across schools testing similar populations of students, as determined by their senior students' average level of parental education and their average freshman CLA+ performance. Effect sizes reflect the standardized difference in performance between freshmen and other class levels tested for the same academic year, and are described in more detail in the "What do my school's effect sizes for CLA+ mean?" section.

When the average performance of seniors at a school is substantially better than expected, this school is said to have high "value added." For instance, consider several schools admitting students with similar levels of parental education and with similar levels of higher-order skills (i.e., freshman CLA+ scores). If, after four years of college education, the seniors at one school perform better on CLA+ than is typical for schools with similar students, one can infer that greater gains in critical-thinking and written-communication skills occurred at the higher performing school. Moreover, one can attribute these greater gains to factors entirely within the school's control, as parental education and freshman CLA+ scores control for differences between students before entering college.

Note that a negative value-added score does not necessarily indicate that no gain occurred between freshman and senior years; rather, it indicates that, according to our model of growth, the observed gain was less than would be expected at schools with similar students.

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. Schools testing more students, and schools with a smaller range of senior CLA+ scores, will typically have narrower confidence intervals, indicating a higher degree of precision of estimation of their value-added score.

Please note that schools must test freshmen in the fall and seniors in the spring of the same academic year to be eligible for value-added scores. If schools have CLA+ data for both groups of students, but have not tested those students in the standard windows, schools can also use the CLA+ value-added model parameters to create their own growth estimates. Instructions are provided on the following page.

What value-added model does CLA+ use?

CLA+ estimates the value added as the difference between freshman and senior deviation scores through an enhanced regression model known as hierarchical linear modeling (HLM), which accounts for CLA+ score variation within and between schools.

Through spring 2009, 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 the HLM approach. Under the HLM 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 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 for the CLA 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).

Originally, CLA+ used seniors' expected academic achievement (EAA) as a predictor instead of seniors' parental education levels. EAA was measured using SAT scores, ACT scores converted into SAT scores, or SLE scores. However, many institutions do not collect standardized testing scores in their admission process. More important, research showed that a much stronger model could be specified by including at least one sociodemographic variable as a predictor. Parental education has been shown to be strongly related to academic achievement in a plethora of studies. Therefore, starting in spring 2017, the CLA+ value-added model will use parental education instead of EAA as both a within-institution and between-institution predictor.

The new HLM value-added model is theoretically stronger than the original HLM value-added model, but it also has equivalent statistical performance on average. For instance, the two sets of value-added scores have very high correlations – on average, at least .80, with much higher results when disattenuation for imperfect reliability is taken into account. Furthermore, split-sample reliability was similar to the values obtained in the previous analysis of original HLM vs. OLS models.

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 CLA+ scores and exiting students with information on their levels of parental education. This information is available in the Student Data File, which is distributed to institutions with each administration's results,
- and the estimated parameters for the value-added model, which are in the appendices to institutional reports.

Using the estimated parameters and the instructions below, one can compute the expected senior CLA+ score for a given school. In combination with the observed mean score for seniors at that school, this can be used to compute the school's value-added score. These values can also be used to perform subgroup analyses or estimate value-added for groups of students that have been tested longitudinally.

Before beginning, parental education must be converted from the "Old Value" to the "New Value" listed in Table 1. The new values reflect average numbers of years of education pertaining to each category, while the old values simply number the categories from 1 to 5 (and use 6 to indicate no information available). Recoding parental education into number of years thus makes the predictor more informative before entering it into the model. Results will not be valid if you skip this step.

Table 1: Recoding Parental Education

Level of education	Old Value	New Value
Less than high school	1	10
High school diploma or equivalent	2	12
Some college but no Bachelor's degree	3	14
Bachelor's degree or equivalent	4	16
At least some graduate education	5	18
Don't know/NA	6	NA

Note. Students who respond "Don't know/NA" to parental education must be removed from the analysis before calculating mean parental education or any mean CLA+ values.

1. Refer to your CLA+ Student Data File to identify your subgroup sample of interest. The subgroup must contain freshmen and seniors with CLA+ scores and seniors with parental education information.
2. Using your CLA+ Student Data File, compute:
    a. The mean level of parental education for 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

Table 2: Estimated Parameters for the Value-Added Model

	Y_{00}	Y_{10}	Y_{01}	Y_{02}	Standard Deviation
Total CLA+ Score	236.2475	6.2791	0.4392	28.8057	43.56
Performance Task	207.628	5.5619	0.4135	31.824	52.50
Selected-Response Questions	265.8832	6.8242	0.4542	26.429	43.71

3. Calculate the expected senior mean CLA+ score using the parameters from the table above. Note that the same equation can be used for the expected senior mean Performance Task score and the expected senior mean Selected-Response Questions score, as long as you change the values of the model parameters *y*s) appropriately.

    Expected senior mean score = Y_{00} + Y_{01}*(Freshman mean CLA+ score) + Y_{02}*(Senior mean parental education) +

4. Use your expected score to calculate your subgroup sample's value-added score:

    Value-added score $_{unstandardized}$ = (Observed senior mean score) – (Expected senior mean score)

5. Convert that value-added score to standard deviation units:

Value-added score $_{standardized}$ = (Value-added score $_{unstandardized}$) / (Standard deviation)

What do my school's CLA+ effect sizes mean?

Effect sizes represent the amount of growth seen from freshman year, in standard deviation units. They are calculated by subtracting the mean freshman performance at your school from the mean sophomore, junior, and/or senior performance, and dividing by the standard deviation of the freshman scores. Effect sizes do not take into account the performance of students at other CLA+ institutions.

Like value-added scores, effect sizes are only provided for institutions that have tested specific class levels in specific windows—freshmen in the fall and seniors in the spring. Schools that have tested students outside of these windows (e.g., those that are assessing students longitudinally) or that want to estimate effect sizes from a different baseline group can easily calculate effect sizes by subtracting the mean performance of one group of students from that of another, and then divide by the standard deviation of the first group's scores.

ANALYSIS

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

The requirements for including students' results in institutional reporting are dependent on the type of report an institution is looking to receive.

For cross-sectional institutional reports, students must:

- test in the correct window, as verified by the registrar (freshmen must test in the fall window and sophomores, juniors, and/or seniors must test in the spring)
- have completed CLA+, which includes submitting a scorable response to the Performance Task, attempting at least half of the Selected-Response Questions, and responding to the CLA+ survey questions

For single-administration institutional reports with cross-CLA+ comparison, students must:

- have a class standing provided by the school's registrar
- have completed CLA+, which includes submitting a scorable response to the Performance Task, attempting at least half of the Selected-Response Questions, and responding to the CLA+ survey questions

For institutional reports with one cohort of students and no cross-CLA+ comparisons, students must:

- have completed CLA+, which includes submitting a scorable response to the Performance Task, attempting at least half of the Selected-Response Questions, and responding to the CLA+ survey questions.

On the student level, total scale scores are computed as the sum of the Performance Task and the Selected-Response Question scores. Only students that have provided scorable responses to both sections will receive a total score and be included in the institutional report. However, section scores and subscores will be provided for all students in the institution's Student Data File, where available.

On the school level, each score is the average of scores from those students that have met the criteria outlined above. Students who have incomplete results are not used in this process. For instance, a student who provides a scorable PT response but does not attempt at least half of the SRQ items will receive a PT score but no SRQ score or total score. This student's PT score will not be used in computing the class-wide mean PT score. Note that, during the registrar data collection process, schools can identify students (e.g., those that are part of an in-depth sample) for exclusion from institutional analyses by assigning them a program code.

Does CLA+ analysis account for ceiling effects?

No school-level scores approach the theoretical maximum of scaled CLA+ scores. There are, however, individual students who have achieved a maximum scale score on the CLA or CLA+, as a function of exceptional performance.

Does CLA+ correct for range restriction?

The CLA+ National Summary Report shows that the demographics and academic abilities of the students who take CLA+ are similar to those of all U.S. college students. Thus, correcting for range restriction is not necessary here. For instance, 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 (fall 2013, n=124) for the same variable, we find a minimum of 752, mean of 1031, maximum of 1354, and standard deviation of 110 (College Results Online, 2010). Information on demographic variables can be obtained from the national summary report.

CORRELATIONS WITH OTHER MEASURES

CLA+ has not been correlated with other measures yet. However, given that CLA+ is essentially the CLA assessment with the addition of SRQs, the following information might be of interest to the reader.

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 were 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 EAA of students?

To date, we have not conducted research to determine whether any linkages or agreements exist between the CLA and various standardized placement tests that would determine an initial freshman course. 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 CLA+?

The reliability of CLA+ scores is assessed from multiple perspectives during each administration.

Performance Tasks are scored through a combination of automated and human scoring. More specifically, each PT is double-scored—once by a machine using Latent Semantics Analysis and once by a trained human scorer. The degree of agreement between scorers is known as the inter-rater reliability or inter-rater correlation. Scores close to 1 indicate high agreement, whereas scores close to 0 indicate little or no agreement. The inter-rater correlation was used as the reliability coefficient for the PT, whereas Cronbach's alpha was utilized for the SRQs. Cronbach's alpha measures the internal consistency of a set of items and can range from 0 to 1. Values closer to 1 indicate higher reliability; values closer to 0 indicate lower reliability. Table 3 shows the reliability statistics for the different components of CLA+.

Table 1: Reliability indices for CLA+ Sections

CLA+ Section	Reliability
Total CLA+	.81
Performance Task	.77
Selected-Response Questions	.76
Scientific & Quantitative Reasoning	.51
Critical Reading & Evaluation	.58
Critique an Argument	.52

Reliability for the PT (*r*= .77) is comparable to the reliability for the SRQs (α=.76). Stratified alpha (Cronbach, Schonemann, & McKie, 1965) was used to combine the PT with the SRQs, resulting in a reliability coefficient of .81. Previous research indicated that CLA scores have been very reliable at the institution level (α =.80) (Klein, et al., 2007), but not at the individual student level (α = .45). However, with CLA+'s addition of SRQs to the exam, the reliability of individual student scores is high enough to ensure the appropriateness of making interpretations at the individual student level and for making inferences in regard to grading, scholarships, admission, or placement.

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 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 skills.

Results from the study show that for critical-thinking skills, the CLA is indeed strongly positively correlated with other tasks that measure such skills. 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, http://www.cae.org/content/pdf/TVS_Report.pdf.

Information on the construct validity of CLA+ will be reported in the near future.

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 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 CLA+ were vetted by a sample of judges who participated in the CLA+ standard-setting study.

After reviewing CLA+ tasks in depth and reading a range of student responses, these judges completed a questionnaire to express their perceptions of the tasks.

As shown in Figure 1, results indicate that the judges perceived the CLA+ tasks to be good assessments of critical-thinking, written-communication, analytic-reasoning, and problem-solving skills. Responding on a 1-5 scale, judges felt, for example, that CLA+ measures important skills that college graduates should possess (Mean 4.92, SD 0.29); students need good analytical-reasoning and problem-solving skills to do well on the task (Mean 4.75, SD 0.45); students need good writing skills to do well on the task (Mean 4.08, .90) and students who do well on the task would also perform well in a job requiring good written-communication (Mean 4.20, SD 0.78), or analytic-reasoning and problem-solving skills (Mean 4.75, SD 0.45). Respondents also agreed, after viewing the tasks, that successful performance on CLA+ may help students compete in a global market (Mean 4.40, SD 0.70).

How are cut scores determined?

On December 12, 2013, a standard-setting study was conducted to formally establish fair and defensible levels of mastery for CLA+. A follow-up study was conducted in December 2014 to establish the cut score for an additional mastery level beyond the four levels determined in the prior year's standard-setting study. The design and execution of the standard-setting studies for CLA+ were consistent with procedures adopted in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999). Relevant practices recommended in these documents were applied to study activities relating to the selection and training of the panel of judges, selection and implementation of the standard-setting methods, provision of feedback to the panel, and documentation of the findings.

CAE recruited a panel of 12 subject matter experts based upon industry standards (Jaeger, 2005). The participants in this study, representing various sectors of both higher and secondary education and content experts, either supervise or work with students/new graduates.

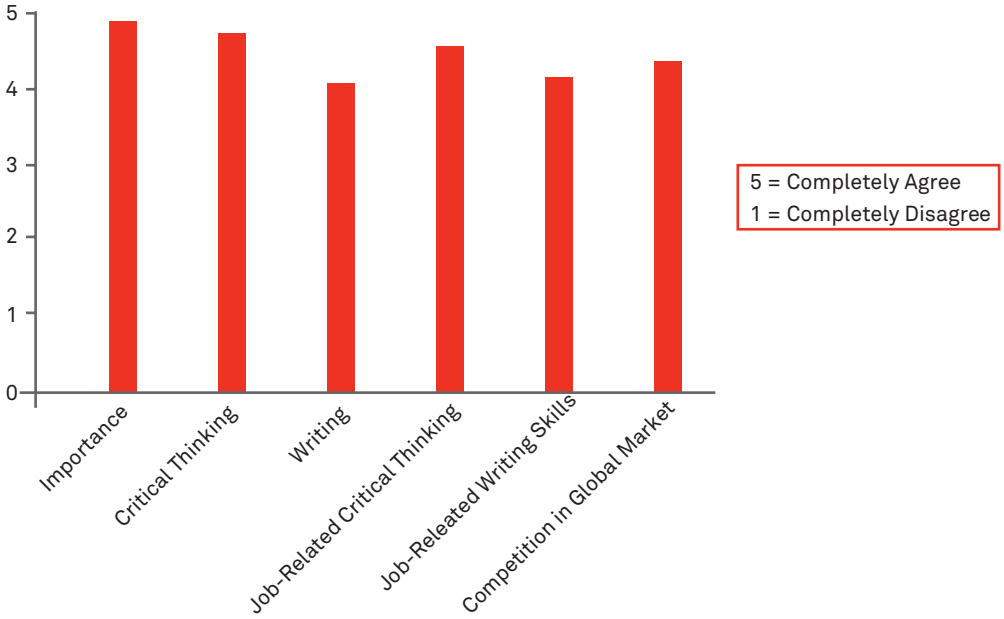

Figure 1: Average face validity assessments of the CLA+

CAE employed the Bookmark (Lewis, Mitzel, Green, & Patz, 1999) methodology to establish the four different cut scores for five different levels of mastery: Below Basic, Basic, Proficient, Accomplished, and Advanced. Under the Bookmark method, the CLA+ SRQ items and PT responses are arranged in order of difficulty and the expert judges on the panel are individually asked to pick the point at which, using the SRQs as an example, a below basic/basic/proficient/accomplished/advanced student would answer this item correctly. So if a judge thought that out of 25 items, a basic student would answer the first seven items correctly, a proficient student would answer the first 14 items correctly, an accomplished student would answer the first 18 items correctly, and an advanced student would answer the first 21 items correctly, their scores would be 7, 14, 18, and 21, respectively. The cut scores for each section and each level of mastery are computed using the average across all 12 panel participants.

STUDENT EFFORT

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

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 EAA, student effort only explains about three to seven percent of the variance in school-level scores (Klein, Benjamin, Shavelson, & Bolus, 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 EAA) motivation was a significant predictor of CLA scores on the student level, but not on the school level (Steedle, 2010b).

Tying stakes to an assessment has also been shown to increase motivation and—in turn—test scores, based on analyses of college students' performance on the ETS Proficiency Profile (Liu,

Bridgeman, & Adler, 2012). Students who were informed that their scores might be shared with faculty at their college or with potential employers performed better than students who were told that their results would be averaged with those of their peers before possibly being shared with external parties. Both of these groups of students performed better than those who were informed that their results would not be seen by anyone outside of the research study.

Because CLA+—unlike its predecessor, the CLA—is reliable at the student level, stakes can be tied to student performance to increase motivation and improve outcomes. To increase motivational opportunities, CAE has embedded results-sharing components into the assessment, delivering electronic badges to students who have performed at or above the proficient level on CLA+, and entering into partnerships will online transcript services and virtual job fairs to allow high-performing students to share their results.

Student Effort and Engagement Survey Responses

Tables 4 and 5 are the summarized results for the questions from the student survey that was administered to participants following the completion of the CLA+ assessment.

Over half the students surveyed (53.6%) put more than a moderate amount of effort into their CLA+ responses and 66.2% of students found the tasks to be moderately to extremely engaging. These results are encouraging because low student motivation and effort are construct-irrelevant threats to the validity of test score interpretations. If students are not motivated, their scores will not be accurate reflections of their maximum level of competency. Although these responses are self-reported, the validity of CLA+ should be enhanced given that stakes are attached to the assessment. Previous research suggests that student motivation and performance is improved as a direct function of attaching stakes to an assessment (Liu, Bridgeman, & Adler, 2012).

Table 4: How much effort did you put into these tasks?

Responses	Percentages
No effort at all	1.6%
A little effort	8.4%
A moderate amount of effort	36.4%
A lot of effort	32.8%
My best effort	20.8%

Table 5: How engaging did you find the tasks?

Responses	Percentages
Not at all engaging	11.3%
Slightly engaging	22.5%
Moderately engaging	39.2%
Very engaging	22.2%
Extremely engaging	4.8%

Face Validity

Students were asked about their perceptions in terms of how well CLA+ measures writing and analytic reasoning and problem solving skills (Table 6).

In an attempt to establish face validity for CLA+, the tasks are designed to emulate critical-thinking and writing tasks that students will encounter in nonacademic endeavors.

Table 6: How well do you think these tasks measure the following skills?

Responses	Writing - PT	Analytic Reasoning and Problem Solving - PT	Analytic Reasoning and Problem Solving - SRQ
Not well at all	5.4%	4.2%	6.8%
Slightly well	17.2%	15.9%	22.3%
Moderately well	45.2%	39.9%	44.0%
Very well	27.3%	31.5%	22.5%
Extremely well	5.0%	8.4%	4.4%

As shown in Table 6, results indicate that the students perceived the tasks to be moderately to extremely good assessments of writing (77.5%) and analytic-reasoning and problem-solving skills (79.8%) for the Performance Tasks, and analytic-reasoning and problem-solving skills (70.9%) for the SRQs.

What is the relationship between CLA+ scores and time spent on CLA+ tasks?

There are moderate positive correlations between CLA+ scores and time spent on CLA+ PTs (Table 7) and between SAT scores and time spent on CLA+ tasks. This relationship is not surprising given that the average test time for tasks in minutes (Table 8) was moderate. Well-developed responses are expected to take longer to compose, although it is possible that students can achieve a high score with a brief response. Table 8 also indicates that students did not invest much time in the SRQs and consequentially, a low correlation is observed between time spent on the SRQs and total scores (Table 7).

Table 7: Relationship between time spent on CLA+ sections and CLA+ total scores

	Time SRQ	Time PT	Total Time	Total Score
Time SRQ	1			
Time PT	.251	1		
Total Time	.594	.928	1	
Total Score	.121	.307	.302	1

Table 8 shows the average testing time for each component of CLA+. Results indicate that on average students finished the different components of the assessment with some time remaining in each section.

Table 8: Test time for tasks in minutes

	Mean	Standard Diviation
Time Spent PT	33.11	14.90
Time Spent SRQ	20.68	6.92
Total Test Time	53.80	17.93

REFERENCES

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

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards of Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

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.

Cronbach, L. J., Schonemann, P., & McKie, D. (1965). *Alpha coefficients for stratified-parallel tests*. Educational and Psychological Measurement, 25, 291-312.

Dorans, N. J., & Schneider, D. (1999). *Concordance between SAT I and ACT scores for individual students* Research Notes, College Entrance Examination Board.

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

Jaeger, R. R. (2005). *Selection of judges for standard setting*. Educational Measurement: Issues and Practice, 10(2), 3-14.

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., Sconing, 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

Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The Bookmark standard setting procedure*. Monterey: McGraw-Hill.

Liu, L., Bridgeman, B., Adler, R. (2012). *Measuring learning outcomes in higher education: motivation matters*. Educational Researcher 2012 41(9), 352-362.

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.