



cla+

TECHNICAL FAQs

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, CLA was designed to measure an institution's contribution, or value added, to the development of these higher-order skills to their student body, and therefore the institution—not the student—was the primary unit of analysis.

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. CLA Performance Tasks presented a real-world situation in which an issue, problem, or conflict is identified, and students are 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 are presented with a statement around which they must construct an argument (Make an Argument), and another in which students are given a logically flawed argument that they must then critique (Critique an Argument).

In its original form, the utility of 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 CLA was restricted to the purposes of its value-added measure; 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 set of selected-response questions, 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 it also affords the opportunity to obtain reliable results at both the institution and student levels. As in its former incarnation, CLA+ allows for an institution to compare its student learning results with the learning results at similarly selective institutions, and 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 their peers at the same class level both within an institution and across CLA+. Unlike the CLA reports, the CLA+ reports will 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 new selected-response questions (SRQs) are also developed with the intent to assess higher-order cognitive skills rather than the recall of factual knowledge. Similar to the Performance Task, 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 newspaper articles.

In addition to the subscores previously reported for the Performance Task (Analysis and Problem Solving, Writing Effectiveness, and Writing Mechanics), CLA+ uses selected-response items to measure students' performance on the following higher-order skills: Scientific and Quantitative

Reasoning, Critical Reading and Evaluation, and Critique an Argument (formerly a constructed-response section of the Analytic Writing Task).

As was the case with the CLA, CLA+ requires 90 minutes of testing time, during which students will have an hour to complete a Performance Task and half an hour to respond to the 25 selected-response questions. Like performance tasks, the selected-response items require students to analyze accompanying documents. Students then answer questions that are designed to measure their ability to apply scientific and quantitative reasoning, critically read and evaluate the texts, and detect logical flaws and questionable assumptions to critique an argument.

Additionally, CLA+ also introduces a new metric in the form of mastery levels. These mastery levels indicate how well students have mastered the higher-order skills measured by CLA+. The mastery level categories are: Below Basic, Basic, Proficient, and Advanced. These levels are determined by specific cut scores established in a standard setting study conducted by CAE with a panel of experts in fall 2013. More information regarding the standard setting study is presented in this document.

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 of subjects such as US 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 a selected-response set. 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, if a student chooses one of these less supported conclusions, he or she can still perform well on the PT if the response aligns to the higher range of the scoring rubric.

The SRQs, like the PTs, represent real-world scenarios and problems. 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 which 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. During piloting, students complete one pilot task as if it were an operational task. 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.

For the selected-response items, 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 vs. 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. The item difficulty, called a "p-value," is the proportion of students answering the item correctly and the total number of students who answered the question. 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 they do not discriminate well between students with low test scores and students with high test scores. 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 results. Because each prompt may have differing possible arguments or relevant information, scorers receive prompt-specific guidance in addition to the scoring rubrics. Additionally, the reported subscores are not adjusted for difficulty like the overall CLA+ scale scores, and therefore are not directly comparable to each other. These PT scores are intended to facilitate criterion-referenced interpretations, as defined by the rubric.

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.

WE 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 diction (word choice and usage).

The CLA+ rubric is available on our website at http://cae.org/images/uploads/pdf/CLA_Plus_Scoring_Rubric.pdf.

The selected-response 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 are CLA+ tasks scored?

During piloting of any new Performance Tasks, all responses are double-scored by human scorers. These double-scored responses are then used to train human scorers and a machine-scoring engine for all operational test administrations.

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 has 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 to increase scoring accuracy, reduce the amount of time between a test administration and report 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 about 800 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.

Once tasks are fully operational, CLA+ uses a combination of automated and human scoring for its Performance Tasks. Each Performance Task response is double-scored: once by IEA, and once by a trained human scorer. Although nearly all responses are scored by both a human scorer and IEA, occasionally IEA identifies unusual responses, which are then automatically sent to the human scoring queue.

All scorer candidates undergo rigorous training in order to become certified CLA+ scorers. Training includes an orientation to the prompts and scoring rubrics/guides, repeated practice grading a wide range of student responses, and extensive feedback and discussion after scoring each response.

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 Verification Papers. The system does not indicate when a scorer has successfully scored a Verification Paper, but 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.

Each response receives subscores in the categories of Analysis & Problem Solving, Writing Effectiveness, and Writing Mechanics. Subscores are assigned on a scale of 1 (lowest) to 6 (highest). Blank responses or responses that are entirely unrelated to the task (e.g., writing about what they had for breakfast) are flagged for removal from results.

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).

How are scorers trained and evaluated?

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.

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.

SCALING PROCESS

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

For each PT, raw subscores are summed to produce a total PT score, which is then scaled to a total score, where PT and SRQ tasks are equally weighted at .50 each. 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 entering academic ability (EAA) as measured by SAT or cross-walked ACT scores of the entering freshmen from the fall 2012 administration of the CLA. 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 a 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 ensures 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. However, unlike the SAT scale, the CLA+ scores do not have a ceiling. This means that if a student achieves an exceptionally high or low raw score, this scaling procedure may produce scale scores outside the normal SAT (Math + Verbal) score range of 400 to 1600.

Do scaling equations change with each administration?

Since 2006, CLA scores have only changed with the introduction of major innovations to the assessment. In fall 2010, CLA equations were revised to reflect the establishment of subscores. The new format of the assessment (CLA+), introduced in fall 2013, has also led to new scaling equations to adapt to the new components of the assessment.

Beginning with the fall 2013 administration of CLA+, scaling equations will remain fixed across administrations, and will only be updated if the population of participating institutions has changed enough to significantly alter score distributions.

Schools wishing to compare CLA+ results to prior year results can use the following equation, which was derived by comparing CLA and CLA+ total scores from the 134 institutions that tested students on both the new and old forms of the assessment ($r=0.822$):

$$\text{CLA scores from fall 2010 – spring 2013: } score_{CLA+} = 184.188 + (0.812 \cdot score_{CLA})$$

$$\text{CLA scores from before fall 2010: } score_{CLA+} = 268.066 + (0.6897 \cdot score_{CLA})$$

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 EAA and freshman CLA+ performance. Effect sizes reflect the standardized difference in performance between freshmen and other class levels tested, and are described in more detail in the “What do my school's effect sizes for the 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 average performance on general academic ability tests (e.g., the SAT or ACT) and on tests of higher-order skills (e.g., CLA+). If, after four years of college education, the seniors at one school perform better on 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 EAA.

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 schools must test freshmen in the fall CLA+ window and seniors in the spring window of the same academic year to be eligible for value-added scores. However, if schools have CLA+ data for both groups of students, but have not tested those students in the standard windows (e.g., schools interested in tracking students longitudinally), 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).

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,
- and the estimated parameters for the value-added model, which will be provided in the appendices to institutional reports starting in spring 2014.

Following the completion of the spring administration of CLA+, the table below will be updated to show the estimated parameters for the CLA+ value-added model. Using these 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.

Estimated Parameters for the Value-Added Model					
	$\gamma_{.00}$	$\gamma_{.10}$	$\gamma_{.01}$	$\gamma_{.02}$	Standard Deviation
Total CLA+ Score	450.47	0.48	0.44	0.20	43.56
Performance Task	442.73	0.39	0.35	0.29	52.50
Selected-Response Questions	454.37	0.57	0.50	0.14	43.71

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 EAA scores.

2. Using your CLA+ Student Data File, compute:

- The mean EAA score of seniors (exiting students) in the sample
- The mean CLA+ score of freshmen (entering students) in the sample
- The mean CLA+ score of seniors (exiting students) in the sample

3. Calculate the senior sample's expected mean CLA+ score, using the parameters from the table above. Please note that the same equation can be used for the individual sections of the CLA+, as well as for the total CLA+ score, by placing the appropriate parameter values from the appropriate row in the table into the equation:

$$\text{Expected Score} = \gamma_{00} + \gamma_{01} (\text{senior mean EAA}) + \gamma_{02} (\text{freshman mean CLA+ score})$$

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

$$\text{Value-added score, unstandardized} = (\text{Observed senior mean score}) - (\text{Expected senior mean score})$$

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

$$\text{Value-added score, standardized} = \frac{(\text{Value-added score, unstandardized})}{(\text{Standard Deviation})}$$

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 of your sophomore, junior, or senior performance, and dividing by the standard deviation of your 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 sophomores, juniors, or 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 dividing 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 upon the type of report an institution is looking to receive.

For cross-sectional (value-added) institutional reports, students must:

- be in the correct class year, as verified by the registrar (freshmen must test in the fall window and sophomores, juniors, or seniors must test in the spring);
- have either an ACT Composite, SAT Math and SAT Verbal/Critical Reading score, or an SLE score;
- and 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;
- and have completed the 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 weighted sum of the Performance Task (weighted at .50) and the Selected-Response Question scores (weighted at .50). Only students that have provided scorable responses to both sections will receive a total score. 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. 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. Originally, CAE capped the distribution of the CLA 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, and CLA+ also does not have a capped distribution.

Does 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 (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).

CONTROLLING FOR EAA

How does 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).

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 selected-response questions, 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 1 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 ($\alpha = .76$). Stratified alpha (Cronbach, Schonemann, & McKie, 1965) was used to combine the PT with the SRQs, resulting in a reliability coefficient of .81. Previous research has indicated that CLA scores have been very reliable at the institution level ($\alpha = .80$) (Klein, et al., 2007), but not at the individual student level ($\alpha = .45$). However, with the addition of SRQs to the PT, 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.

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

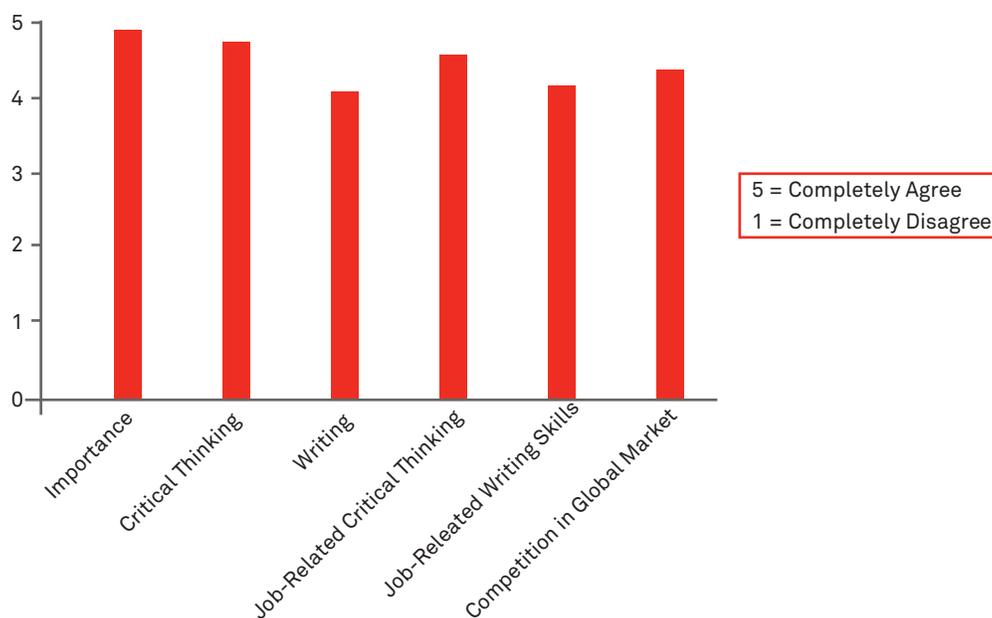


Figure 1: Average face validity assessments of the CLA+

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+. The design and execution of the standard setting study 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.

CAE employed the bookmark (Lewis, Mitzel, Green, & Patz, 1999) methodology to establish the three different cut scores for four different levels of mastery: Below Basic, Basic, Proficient, 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 basic/proficient/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, and an advanced student would answer the first 21 items correctly, their scores would be 7, 14, and 21. The overall cut scores for each section and each level of mastery is 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 with online transcript services and virtual job fairs to allow high-performing students to share their results.

Student Effort and Engagement Survey Responses

Tables 2 and 3 are the summarized results for the questions from the student survey that was administered to participants following the completion of the CLA+ assessment.

Tables 2 and 3 show that 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 1: Reliability indices for CLA+ Sections

How much effort did you put into these tasks?	
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 3: Engaging

How engaging did you find the tasks?	
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 4).

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.

As shown in Table 4, 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.

Table 4: Face Validity

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

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+ PTs (Table 5) 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 6) 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 6 also indicates that students did not invest much time in the SRQs and consequentially, a low correlation is observed between the time spent on SRQs and the total score (Table 5).

Table 5: 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 6 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 6: 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., Scoring, J., Bolus, R., Bridgeman, B., Kugelmass, H., . . . Steedle, J. (2009). *Test Validity Study (TVS) Report*. Supported by the Fund for the Improvement of Postsecondary Education. from http://www.cae.org/content/pdf/TVS_Report.pdf
- 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.