CAE

CLA+

Frequently Asked Questions

www.cae.org

1. What is the Collegiate Learning Assessment?

The Collegiate Learning Assessment (CLA+) is a performance-based assessment of critical thinking and written communication. It includes two components: a Performance Task (PT) and a set of Selected-Response Questions (SRQs). The PT presents students with a real-world situation that requires a written response. Students are asked to address an issue, propose the solution to a problem, or recommend a course of action to resolve a conflict. They are instructed to support their responses by utilizing information provided in a Document Library, which contains a variety of reference materials, such as technical reports, data tables, newspaper articles, office memoranda, and emails.

The SRQs are aligned to the same construct as the PT and measure students' critical-thinking skills. Like the PT, the SRQs are document based and require students to draw information from the materials provided to answer the questions. The supporting documents may include letters, memos, photographs, charts, and newspaper articles. Students are expected to answer questions that require them to use data literacy skills (10 questions), read critically to evaluate a situation (10 questions), and identify logical fallacies in an argument (5 questions). These types of questions require students to think at a deeper level than the traditional recall-and-recognition questions.

2. How long does the CLA+ take to administer?

The CLA+ requires 90 minutes of testing time. Students have 60 minutes to complete the PT and 30 minutes to complete the SRQs.

3. When was the CLA+ developed and for what purpose?

Prior to the development of the CLA+, the CLA was administered to measure an institution's contribution, or value added, to the development of higher-order skills to its student body. The CLA employed a matrix sampling approach, in which students were randomly assigned either a Performance Task or an Analytic Writing Task. The institution, rather than the student, was the unit of analysis.

In the Fall of 2013, the CLA+ was launched. Students take both components of the test: PT and SRQs. The test design affords the opportunity to obtain reliable results at the institution and student levels. The CLA+ results at the student level allow for a comparison of how well a student performed relative to his or her peers at the same class level both within an institution and across the other CLA+ institutions.

© 2022 Council for Aid to Education, Inc.

4. What is the sample of CLA+ test takers?

To date, approximately 200,000 US students and 25,000 international students have taken the CLA+. The number of participating institutions in the United States is over 350.

Compared to the national rate, the group of higher education institutions administering the CLA+ tends to have a higher representation of public schools (60% vs. 30% nationally) and Hispanic-Serving Institutions (18% vs. 10%) and a slightly lower representation of Historically Black Colleges and Universities (3% vs. 4%).

Based on the Carnegie Classification, the group of colleges participating in the CLA+ has a slightly lower percentage of schools granting doctoral degrees (15% vs. 17% nationally), a higher percentage of schools granting master's degrees (55% vs. 39%), and a lower percentage of schools granting baccalaureate degrees (30% vs. 45%). Barron's Admissions Competitiveness Index indicates a lower mean score compared to the national average (3.1 vs. 3.6).

The population of students has the following characteristics:

- The majority of students are White (55%), followed by Hispanic/Latino (13%), African American/Black (13%), Asian (10%), other (3%), American Indian/Alaska Native/Indigenous (< 1%), and Native Hawaiian/other Pacific Islander (< 1%). A small percentage declined to state (4%).
- A higher percentage of students are female compared to male (57% vs. 40%).
- English is the primary language (84%).
- About half of the students (52%) reported at least one parent or guardian had attained a college degree or higher.
- Students' fields of study were Sciences and Engineering (26%), Helping/Services (24%), Business (17%), Social Sciences (12%), and Humanities and Languages (11%).

5. How are CLA+ PTs and SRQs developed?

CAE test developers follow rigorous and structured task and item development plans when creating the PTs and SRQs. The primary goal is to develop assessment tasks and items that are authentic and engaging for students. The documents for both the PTs and the SRQs are presented in the most appropriate format for the scenario (e.g., abstract from a journal, graph, memo, blog post, newspaper article, report).

The PTs are created so that students can craft an argument using only the information provided and there is enough information to support or refute a position from multiple perspectives. The SRQs are created to measure the content intended. After several rounds of revisions between the developer and one or more of CAE's editors, PTs and SRQs are selected for pilot testing.

The purpose of pilot testing is to determine whether revisions are needed before the tasks and items are used on an operational test. For the PT, scoring procedures are evaluated and final revisions are made to the tasks to ensure that the task is eliciting the types of responses intended. For the SRQs, statistical analyses are performed to ensure that the items are within the expected difficulty range and the items correlate with the test score, indicating that the items measure the same construct.

© 2022 Council for Aid to Education, Inc.

6. What scores are reported for the CLA+?

Students receive multiple pieces of information about their performance on the CLA+. The mastery level is based on the CLA+ total score. The five mastery levels, in order of progression, are Emerging, Developing, Proficient, Accomplished, and Advanced.

Students receive an overall score for the PT section and three subscores for Analysis and Problem Solving (APS), Writing Effectiveness (WE), and Writing Mechanics (WM). The subscores are based on a scoring rubric in which 1 is the lowest level of performance and 6 is the highest level of performance.

APS measures a student's ability to make a logical decision or conclusion and support it with accurate and relevant information from the Document Library. The skills measured include the following:

- Identifying facts or ideas and interpreting them accurately
- Identifying connected and conflicting information
- Analyzing logic and identifying assumptions in arguments
- Evaluating the reliability of information
- Deciding on a course of action to solve a problem

WE assesses a student's ability to construct and organize logically cohesive arguments to support a position. The skills measured include the following:

- Stating a position clearly
- Presenting evidence in support of an argument
- Elaborating on facts or ideas
- Constructing an organized and logically cohesive argument
- Including the use of effective transitions

WM evaluates a student's facility with the conventions of standard written English and use of language and vocabulary. The skills measured include the following:

- Using vocabulary correctly
- Demonstrating effective use of varied and complex vocabulary
- Constructing grammatically and syntactically correct sentences
- Varying the structure and complexity of sentences

Students receive an overall score for the SRQ section and three subscores for the SRQ subsections: Data Literacy (DL), Critical Reading and Evaluation (CRE), and Critiquing an Argument (CA). The SRQ section score and the SRQ subscores are reported as scale scores. The skills measured by the subscores are as follows.

© 2022 Council for Aid to Education, Inc.

Data Literacy

- Making inferences and hypotheses
- Evaluating data collection methods
- Detecting questionable assumptions
- Supporting or refuting a position with scientific evidence
- Drawing a conclusion
- Recognizing when additional research is required

Critical Reading and Evaluation

- Supporting or refuting a position
- Analyzing logic
- Identifying assumptions in arguments
- Evaluating the reliability of information
- Identifying connected and conflicting information

Critiquing an Argument

- Detecting logical flaws and questionable assumptions
- Addressing information that could strengthen or weaken an argument
- Evaluating alternative conclusions

7. What is the procedure for calculating the CLA+ scale scores?

CAE performs an equating procedure to convert the sum of the item scores (raw scores) to scale scores. Equating is used in situations in which there are multiple forms of a test and scores earned on different tests are compared to each other (Kolen & Brennan, 1995). Equating ensures that scale scores can be compared with each other regardless of when the test was taken and which PT and set of SRQs were administered.

The equating procedure that CAE uses is linear equating. The result of linear equating is a set of parameters (slope, intercept) that transform each raw score to a scale score. The scale score average and standard deviation of the norm population and the raw score average and standard deviation for a given PT or set of SRQs are used in the computation. An additional step is performed to compute the weighted average of the SRQ subscores to generate scale scores for the SRQ section. The PT and SRQ scale scores are averaged together to create the CLA+ total score. Periodically, CAE updates the norm population information and the equating equations to ensure that students receive valid scores.

© 2022 Council for Aid to Education, Inc.

8. How do I interpret my school's growth estimates on the CLA+?

CLA+ includes two forms of growth estimates—value-added scores and effect sizes. Schools that test freshmen in the fall and seniors in the spring of the same academic year are eligible for value-added scores. For schools that want to conduct their own value-added analyses, CAE will provide the necessary information. For example, schools may want to compute their own growth estimates for students that test outside of the standard windows or for a subgroup of students.

Value-added results

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 performance on the CLA+.

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 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. A negative value-added score indicates that the observed gain was less than would be expected at schools with similar students. It does not indicate that no gain occurred between freshman and senior years. Value-added scores are placed on a normalized (z-score) scale and assigned performance levels, as shown in Table 1.

Table 1. Value-Added Performance Levels

Score	Performance level
< -2.0	Well below expected
-2.0 to -1.0	Below expected
-1.0 to 1.0	Near expected
1.0 to 2.0	Above expected
> 2.0	Well above expected

© 2022 Council for Aid to Education, Inc.

Effect size results

Effect sizes reflect the standardized difference in performance between freshmen and other class levels tested in the same academic year for a given institution. Effect sizes represent the amount of growth seen from freshman year, in standard deviation units. They are calculated by subtracting the average freshman performance at a given school from the average sophomore, junior, or senior performance, and dividing by the standard deviation of the freshman scores.

A positive effect size indicates that the more advanced students (e.g., sophomore, junior, senior) performed better than the freshmen class overall, whereas a negative effect size indicates that the freshmen students performed better than the comparison class overall. The magnitude of the effect size reflects the difference in mean scores. For example, an effect size of 1.0 indicates that on average the senior students performed 1.0 standard deviation higher than the freshmen students. A small effect size (< 0.20) indicates limited practical significance.

9. What is the reliability of CLA+?

Reliability of a test refers to the consistency of students' test scores on parallel forms of a test. A reliable test is one that produces scores that are expected to be relatively stable if the test is administered repeatedly under similar conditions. Because it is impractical to administer multiple forms of a test, reliability is estimated on a single administration of the test. This type of reliability, known as internal consistency, provides an estimate of how consistently examinees perform across test items during a single test administration (Crocker & Algina, 1986).

The internal consistency of a set of items ranges from 0 to 1, with higher values indicating higher reliability. Internal consistency estimates typically are lower for sets with fewer items. On the CLA+, the internal consistency estimates for the subscores can be noticeably lower than those for the section and the full test scores and this should be taken into consideration when interpreting scores at the subscore level.

For scores that include single-point items (i.e., multiple-choice), Cronbach's (1951) coefficient alpha is used to estimate reliability. For scores that include a combination of single- and multiple-point items, the stratified coefficient alpha (Cronbach et al., 1965) is used to estimate reliability.

According to the *Standards for Educational and Psychological Testing*, "the need for precision of measurement increases as the consequences of decisions and interpretations grow in importance" (American Educational Research Association [AERA] et al., 2014, p. 33). Although the *Standards* document does not include guidelines, as a rule of thumb reliability coefficients that are 0.80 or higher are considered acceptable for tests of moderate lengths. Table 2 shows that the average reliabilities across the operational test forms are within a reasonable range to report scores.

© 2022 Council for Aid to Education, Inc.

Table 2. Reliability of the CLA+ Test Scores

Score	Points	Average	Minimum	Maximum
SRQ section	25	0.79	0.78	0.80
DL	10	0.59	0.56	0.61
CRE	10	0.64	0.53	0.69
CA	5	0.49	0.41	0.57
PT section (3 rating scales)	18	0.92	0.90	0.94
Total test	43	0.85	0.85	0.87

Note: Stratified alpha results are presented for the total test score, and Cronbach's alpha is reported for all other scores.

Performance Task scores

PT scores are generated by two rater scores, either a combination of artificial intelligence (AI) machine scoring and human scoring or by two human raters. The degree of agreement between scorers is known as the inter-rater reliability. CAE summarizes the reliability of the PT scores by presenting the percentage of agreement of rater scores for exact and exact-plus-adjacent scores. In addition, the kappa statistic is calculated to reflect the level of improvement beyond chance in the consistency of scoring. Table 3 presents the rater agreement for the PT scores. To aid in the interpretation of the kappa statistic, the classifications proposed by Altman (1991) and Landis and Koch (1977) are presented in Table 4.

The percentage of rater agreement (i.e., exact, exact-plus-adjacent scores) shows that the sets of raters (AI, human; two human) have comparable results. Generally, exact agreement exceeded 60 percent and exact-plus-adjacent agreement approached 100 percent across the three PT subscores (APS, WE, WM) for the PTs administered. The kappa results show that the strength of rater agreement is good to very good.

© 2022 Council for Aid to Education, Inc.

Table 3. Rater Agreement for PT Subscores

Rater pair	Rater agreement	Average	Minimum	Maximum
AI, human	Percentage exact	64.9	56.5	77.4
	Percentage exact + adjacent	97.5	95.2	99.5
	Weighted kappa	0.81	0.71	0.87
Two human	Percentage exact	61.6	58.4	65.1
	Percentage exact + adjacent	99.0	98.4	99.9
	Weighted kappa	0.84	0.79	0.90

Table 4. Kappa Interpretation

Kappa	Strength of agreement
0	None
< 0.20	Poor
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Good
0.81 to 1.00	Very good

10. What evidence of validity is there for the CLA+?

The purpose of test score validation is not to validate the test itself, but to validate interpretations of the test scores for particular purposes or actions. The results of the CLA+ are used to make inferences about students' critical-thinking and written-communication skills.

Test score validation is an ongoing process, beginning at the initial conceptualization and continuing throughout the entire assessment process. Every aspect of an assessment provides evidence in support of (or challenge to) the validity of an intended interpretation of test scores, including test design, content specifications, item development, psychometric quality, and inferences made from the results.

The *Standards* document gives a framework for describing sources of evidence that should be considered when constructing a validity argument (AERA et al., 2014, pp. 26–31). The sources of validity evidence include evidence based on 1) test content; 2) response processes; 3) internal test structure; 4) relations to other variables; and 5) consequences of testing.

© 2022 Council for Aid to Education, Inc.

Validity evidence based on internal test structure

CAE gathers evidence for internal test structure that includes test reliability, the evaluation of item performance, and the evaluation of the correlation coefficients that measure the relationship between the content strand scores (i.e., subscores).

Evaluation of test reliability

The evaluation of the test reliability indicates that the test and section scores have adequate reliability to support making inferences about students' critical-thinking and written-communication skills. Refer to the section above (What is the reliability of the CLA+?) for the results of reliability studies.

Evaluation of item performance

The average raw scores indicate whether the test is of appropriate difficulty for the student population. Table 5 shows that across administrations, students were able to achieve more than half of the score points on average (see the "Average % of points earned" column). These results suggest that the test and test components are at the appropriate level of difficulty for the student cohort.

Table 5. Average Test Scores

Score	Points	Average raw score	Average % of points earned
SRQ section	25	14.15	56.6%
DL	10	5.40	54.0%
CRE	10	6.00	60.0%
CA	5	2.73	54.7%
PT section	18	18	53.2%
Total test	43	43	55.9%

Evaluation of correlation coefficients

The relationships between the test components indicate the degree to which the components correspond to the construct on which the score interpretations are based. The inter-correlations of the SRQ subscores are roughly between 0.40 and 0.50. Similarly, the correlations between the PT and SRQ section scale scores are between 0.40 and 0.50. High correlations were observed between the CLA+ total scale score and the PT scale score (r = 0.82 to 0.83) and the CLA+ total scale score and the SRQ scale score (r = 0.85 to 0.87).

© 2022 Council for Aid to Education, Inc.

Validity evidence based on response processes

CAE collects evidence based on response processes. Students indicate their level of effort and engagement on the test sections in the student survey following the CLA+. A 5-point Likert rating scale is used to convey their effort (1 = no effort at all; 5 = my best effort) and engagement (1 = not at all engaging; 5 = extremely engaging) on the PT and SRQ sections.

The results of student opinion questions, shown in Table 6, reveal that the average scores are above the midpoint of the scale (3 out of 5) for all opinion questions except the SRQ section engagement. These findings suggest that students took the assessment seriously, devoted the level of effort one would expect, and were engaged in the tasks presented.

Table 6. Student Effort and Engagement

Opinion	Average	Minimum	Maximum
Effort on PT	3.68	1	5
Effort on SRQs	3.49	1	5
PT engagement	3.19	1	5
SRQs engagement	2.86	1	5

References

Altman, D. G. (1991). *Practical statistics for medical research*. Chapman and Hall.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Wadsworth Group/ Thomson Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J., Schonemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, 25(2), 291–312.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. Springer-Verlag.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

© 2022 Council for Aid to Education, Inc.