cwra+

TECHNICAL FAQs

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

CWRA+ TASKS

How are CWRA+ 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 to 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 and refute from multiple perspectives. One of the unique features of the CWRA+ 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 article, 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 SRQ section 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, 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 real-world scenario or 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 traditional 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 CWRA+ tasks have varied backgrounds including history, English, mathematics, psychology, and psychometrics. Over the years of developing the CWRA and CWRA+, the team now has extensive experience with test development and writing evaluation.

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

Prior to the introduction of the CWRA+, the assessment was only valid and reliable at the institutional level. CWRA+ 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 CWRA 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 spent testing, CAE decided to develop the CWRA+ 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 CWRA.

Additionally, both performance tasks and selected-response questions 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 CWRA+ scoring rubrics?

The CWRA+ 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.

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 CWRA+ rubric is available on our website at http://cae.org/images/uploads/pdf/CWRA_Plus_Scoring_Rubric.pdf.

The SRQ section of the CWRA+ 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 CWRA+ scoring work?

All scorer candidates undergo rigorous training in order to become certified CWRA+ 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 text, not just writing mechanics. Pearson designed IEA for CWRA+ using a broad range of real CWRA+ responses and scores to ensure its consistency with scores generated by human scorers. Thus, human scorers remain the basis for scoring the CWRA+ 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 CWRA+ 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, CWRA+ uses a combination of automated and human scoring for its PTs. 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 CLA Performance Tasks to compare the accuracy of human versus automated scoring. For 12 of the thirteen tasks examined, IEA scores agreed more often with the average of multiple experts (r = .84-.93) than two experts agreed with each other (r = .80-.88). These results suggest that computer-assisted scoring is as accurate as—and in some cases, more accurate than—expert human scorers (Elliot, 2011). We do not have data specific to the CWRA on the accuracy of human versus automated scoring; however, the CLA Performance Tasks are substantially similar to the CWRA Performance Tasks, and are developed by the same group of item writers and editors. We expect machine scoring accuracy would be nearly equivalent between CLA and CWRA, but are working on empirical research to support this contention.

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 CWRA+ total score. The raw total scores undergo a final linear transformation to become scaled CWRA+ 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.

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:

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

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

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

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

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

On the student level, total scale scores are computed as the average 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 on 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 CWRA+ analysis account for ceiling effects?

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

RELIABILITY

What is the reliability of the CWRA+?

The reliability of CWRA+ 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 the CWRA+.

Table 1: Reliability indices for CWRA+ Sections

CWRA+ Section	Reliability
Total CWRA+	.84
Performance Task	.78
Selected-Response Questions	.76
Scientific & Quantitative Reasoning	.51
Critical Reading & Evaluation	.58
Critique an Argument	.52

Reliability for the PT (*r*= .78) 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 .84. Previous research has indicated that CLA and CWRA 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 CWRA+'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 by the pattern of correlations between a test and other tests of similar and different skills (Campbell, 1959). For example, if the CLA measures critical-thinking skills, then it should be highly (positively) correlated with other tasks that measure critical-thinking 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 our website, http://www.cae.org/content/pdf/TVS_Report.pdf.

As noted, prior construct validity information was only available on the CLA. Information on the construct validity of CWRA+ 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 CWRA+ to have face validity, its tasks must emulate the critical thinking and writing challenges that students will face outside the classroom. These characteristics of the CWRA+ were vetted by a sample of judges who participated in the CWRA+ standard-setting study.

After reviewing the CWRA+ 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 CWRA+ 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 the CWRA+ measures important skills that high school graduates should possess (Mean 5.00, SD 0); students need good analytical reasoning and problem solving skills to do well on the task (Mean 4.93, SD 0.26); students need good writing skills to do well on the task (Mean 4.20, .56), and students who do well on the task would also perform well in a job requiring good written communication (Mean 4.67, SD 0.49) or analytic reasoning and problem solving skills (Mean 4.63, SD 0.48). Respondents also agreed, after viewing the tasks, that successful performance on the CWRA+ may help students compete in a global market (Mean 4.58, SD 0.52).

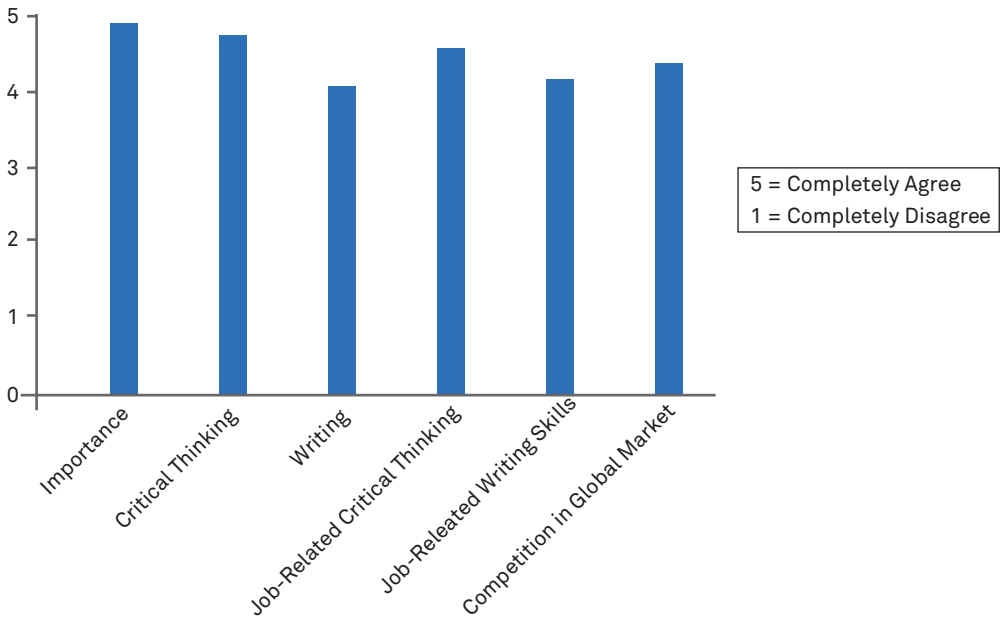

Figure 1: Average face validity assessments of the CWRA+

How are cut scores determined?

On December 13, 2013, a standard-setting study was conducted to formally establish fair and defensible levels of mastery for CWRA+. The design and execution of the standard-setting study for CWRA+ were consistent with procedures adopted in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education). 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 15 subject-matter experts based on industry standards (Jaeger, 2005). The participants on this study, representing various sectors of both higher education and employers, were all content experts who 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 CWRA+ 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 questions 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 15 panel participants.

STUDENT EFFORT

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

CWRA+ does not control for self-reported student effort, but has conducted some research on the role that motivation plays in CWRA achievement. Analyses of the relationship between Performance Task scores and self-reported effort suggest that, controlling for entering academic ability, student effort only explains about three to seven percent of the variance in school-level scores (Klein et al., 2007). Additionally, internal analyses run on all CWRA+ data from Fall 2013 through Spring 2016 show that self-reported effort only accounted for about 6% of the variance in individual-level CWRA+ total scores.

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

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 CWRA+—unlike its predecessor, the CWRA—is reliable at the student level, stakes can be tied to student performance to increase motivation and improve outcomes. To increase motivational opportunities, CAE will soon begin embedding results-sharing components into the assessment, delivering electronic badges to students who have performed at or above the proficient level on CWRA+, and entering into partnerships with online transcript services 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 CWRA+ assessment.

Tables 2 and 3 show that 93.2% put more than moderate amount of effort into their CWRA+ responses and 67.3% 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 the CWRA+ 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 2: Effort

How much effort did you put into these tasks?	
No effort at all	1.2%
A little effort	5.6%
A moderate amount of effort	30.8%
A lot of effort	37.6%
My best effort	24.8%

Table 3: Engaging

How engaging did you find the tasks?	
Not at all engaging	10.8%
Slightly engaging	21.9%
Moderately engaging	38.8%
Very engaging	23.2%
Extremely engaging	5.3%

Face Validity

Students were asked about their perceptions in terms of how well the CWRA+ measures writing and analytic reasoning and problem solving skills (Table 4).

In an attempt to establish face validity for CWRA+, 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 (83.1%) and analytic reasoning and problem solving skills (83.6%) for the Performance Tasks, and analytic reasoning and problem solving skills (72.9%) 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	4.0%	3.5%	6.2%
Slightly well	13.0%	13.0%	20.9%
Moderately well	43.8%	39.1%	44.7%
Very well	33.1%	33.8%	23.7%
Extremely well	6.2%	10.7%	4.5%

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

There are moderate positive correlations between CWRA+ scores and time spent on CWRA+ PTs (Table 5). 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 CWRA+ sections and CWRA+ Total Scores

	Time SRQ	Time PT	Total Time	Total Score
Time SRQ	1			
Time PT	.145	1		
Total Time	.501	.929	1	
Total Score	.232	.370	.411	1

Table 6 shows the average testing time for each component of CWRA+. 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	SD
Time Spent PT	37.81	14.44
Time Spent SRQ	22.27	6.16
Total Test Time	60.08	16.44

REFERENCES

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.

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. (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.