CAE

Test Reliability (Consistency)

What Is Test Reliability?

Reliability of test scores refers to the consistency of students' scores. A reliable test is one that produces scores that are expected to be relatively stable, or consistent, if the test is administered repeatedly under similar conditions. Because it is not feasible to administer a test to students multiple times, reliability can be estimated on a single administration of a test.

The reliability measure called *internal consistency* provides an estimate of how consistently students perform across items within a test during a single test administration (Crocker & Algina, 1986). The closer the reliability coefficient is to 1, which conveys a perfectly consistent test, the more *reliable* the scores are.

According to the *Standards for Educational and Psychological Testing* (American Education Research Association et al., 2014), reliability is always important, but the need for precision increases when the consequences of decisions and interpretations grow in importance.

How Is Test Reliability Used?

Tests assess students' knowledge, skills, and abilities. Tests that are used to make important decisions about students, such as promotion to the next grade or entry into professional practice (e.g., accounting, dentistry) should have high reliability. These tests tend to be longer, because they assess a broad range of content, and yield high reliability (0.90 or higher).

In contrast, formative assessments that are used to gauge students' performance toward learning goals, such as classroom unit tests or quizzes, tend to be shorter, because they assess a narrow range of content, and have lower reliability.

Generally, reliability is lower for tests or subscores that are measured by a small number of items and higher for tests or subscores that are measured by a larger number of items. *Interpretations of test scores should take reliability into account.*

How Should I Use CAE Test Scores?

The CAE reports provide information on the total test scores, section scores, and section subscores. Through the use of aggregate student scores, the Total scores, the Performance Task (PT) and Selected-Response (SR) section scores, and the subscores may be used to make decisions at the institution level. However, only the Total, PT, and SR scores have sufficient reliability indices to be appropriate for decisions regarding students.

As outlined in Table 1, CAE's reliability indices are higher for the SR and PT section scores than their respective subscores. The reliability for the SR section scores approaches or exceeds 0.80, whereas the reliability for the PT section scores is 0.90 or higher. Reliability of the total test scores, based on the stratified alpha model, which combines the reliability of test sections with different score point values, is between 0.86 and 0.91 across the tests. Reliability details for CLA+, CCRA+ (High School), and CCRA+ (Middle School) are included in the Appendix.

Table 1

CAE Reliability Indices

Section/Subscore	CLA+	CCRA+ (HS)	CCRA+ (MS)
Selected-Response (SR) Section	**0.79 to 0.80**	**0.79 to 0.86**	**0.76 to 0.79**
Data Literacy	0.45 to 0.71	0.47 to 0.73	0.47 to 0.67
Critical Reading and Evaluation	0.36 to 0.72	0.42 to 0.72	0.42 to 0.63
Critiquing an Argument	0.32 to 0.63	0.43 to 0.67	0.45 to 0.65
Performance Task (PT) Section	**0.90 to 0.92**	**0.92 to 0.93**	**0.91**
Analysis and Problem Solving	0.41 to 0.65	0.41 to 0.65	0.45 to 0.60
Writing Effectiveness	0.33 to 0.67	0.33 to 0.67	0.47 to 0.61
Writing Mechanics	0.43 to 0.67	0.43 to 0.67	0.51 to 0.67
Total Score	**0.87 to 0.91**	**0.87 to 0.91**	**0.86 to 0.87**

References

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

Brennan, R. l. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement, 32*(4), 385–396.

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

Appendix

CLA+ Reliability Indices

Section/Subscore	Items	Points	Reliability			Decisions	
			Values	Index	Data source	Student	Institution[1]
Selected-Response Section	25	25	0.79 to 0.80	Alpha	Spring21	X	X
Data Literacy	10	10	0.45 to 0.71	Alpha	Fall13—Spring21		X
Critical Reading and Evaluation	10	10	0.36 to 0.72	Alpha	Fall13—Spring21		X
Critiquing an Argument	5	5	0.32 to 0.63	Alpha	Fall13—Spring21		X
Performance Task Section	3	18	0.90 to 0.92	Alpha	Spring21	X	X
Analysis and Problem Solving	1	6	0.41 to 0.65	Exact agreement	Fall13—Spring21		X
Writing Effectiveness	1	6	0.33 to 0.67	Exact agreement	Fall13—Spring21		X
Writing Mechanics	1	6	0.43 to 0.67	Exact agreement	Fall13—Spring21		X
Total Score	28	43	0.87 to 0.91	Stratified alpha	Spring21	X	X

CAE

CCRA+ (HS) Reliability Indices

Subscore/ Test	Items	Points	Reliability			Decisions	
			Values	Index	Data source	Student	Institution[1]
Selected-Response Section	25	25	0.79 to 0.86	Alpha	Spring21	X	X
Data Literacy	10	10	0.47 to 0.73	Alpha	Fall13—Spring21		X
Critical Reading and Evaluation	10	10	0.42 to 0.72	Alpha	Fall13—Spring21		X
Critiquing an Argument	5	5	0.43 to 0.67	Alpha	Fall13—Spring21		X
Performance Task Section	3	18	0.92 to 0.93	Alpha	Spring21	X	X
Analysis and Problem Solving	1	6	0.41 to 0.65	Exact agreement	Fall13—Spring21		X
Writing Effectiveness	1	6	0.33 to 0.67	Exact agreement	Fall13—Spring21		X
Writing Mechanics	1	6	0.43 to 0.67	Exact agreement	Fall13—Spring21		X
Total Score	28	43	0.87 to 0.91	Stratified alpha	Spring21	X	X

CCRA+ (MS) Reliability Indices

Subscore/ Test	Items	Points	Reliability			Decisions	
			Values	Index	Data source	Student	Institution[1]
Selected-Response Section	25	25	0.76 to 0.79	Alpha	Spring21	X	X
Data Literacy	10	10	0.47 to 0.67	Alpha	Fall13—Spring21		X
Critical Reading and Evaluation	10	10	0.42 to 0.63	Alpha	Fall13—Spring21		X
Critiquing an Argument	5	5	0.45 to 0.65	Alpha	Fall13—Spring21		X
Performance Task Section	3	18	0.91	Alpha	Spring21	X	X
Analysis and Problem Solving	1	6	0.45 to 0.60	Exact agreement	Fall13—Spring21		X
Writing Effectiveness	1	6	0.47 to 0.61	Exact agreement	Fall13—Spring21		X
Writing Mechanics	1	6	0.51 to 0.67	Exact agreement	Fall13—Spring21		X
Total Score	28	43	0.86 to 0.87	Stratified alpha	Spring21	X	X

[1] Research has shown that scores may become more reliable when aggregated at the institution level (Brennan, 1995). Aggregated student results may be used to make low stakes and high stakes decisions.