CAE

International Testing of a Performance-Based Assessment in Higher Education

June 7, 2017

Doris Zahner
Fiorella Kostoris

Paper presented at the 2016 American Educational Research Association Annual Meeting, Washington, DC.

Objectives and theoretical framework

International assessments in higher education are especially challenging because differences across countries (e.g., educational systems, SES) increase the complexity of testing (Blömeke, Zlatkin-Troitschanskaia, Kuhn, & Fege, 2013). This becomes even more challenging when using performance-based assessments, which are becoming more prominent in assessment programs (Kahl, 2008; Penfield & Lam, 2000).

ANVUR collaborated with the Council for Aid to Education (CAE) to adapt, translate, administer, and score the CLA+, a performance-based assessment of critical-thinking and written-communication skills, to Italian university students. The objectives of this study were to see if it was feasible to assess Italian students' skills, to conduct a cross-country comparison of Italians and Americans, and to validate the importance of these skills in the labor market.

Method

In 2015, students at participating institutions completed a translated and adapted version of the CLA+ that included the Performance Task (PT) "Life Expectancy" and one of two sets of 25 selected-response questions (SRQs). The CLA+ for the ANVUR project was administered between May and July 2015, and a total of 6,268 students across 23 institutions participated in the administration.

The translated and adapted versions of the CLA+ Performance Tasks were scored in Italy by a team of trained scorers. CAE representatives led a series of training sessions, both virtually via WebEx and on-site in Rome, between January and July 2015.

- ANVUR identified lead scorers from each participating institution.
- CAE trained lead scorers at a two-day session in Rome.
- ANVUR lead scorers identified benchmark papers and trained Italy-based scorers.
- Italy-based scorers completed the remaining scoring.
- CAE provided additional support to lead scorers.

Out of the 6,268 assessments administered, the CLA+ was completed by 6,245 students across 23 institutions. The students scoring 0 on the PTs (n = 23) were removed from the analyses.

Data sources and materials

A translated and adapted version of CLA+ was used in the study. Student responses are measured on three subscales: analysis and problem-solving, writing effectiveness, and writing mechanics. The SRQs measure students' analysis and problem-solving skills on three subscales: Scientific and Quantitative Reasoning, Critical Reading and Evaluation, and Critique an Argument. Students were given 60 minutes for the PT and 30 minutes for the SRQs.

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

Results

Table 1 contains the correlation coefficients for the PT subscores (Analysis and Problem Solving, Writing Effectiveness, and Writing Mechanics) and total score and the SRQ total score. The correlations between the SRQ total and the PT subscores and total score are unusually low (r = .29). This is illustrated in Figure 1. Typically, the correlation between PT and SRQ total scores is at least r = .50. The correlations across the PT subscores are as expected. Similarly low correlations were observed in the 2013 data set (r = .23–.28) and ANVUR and CAE concluded that perhaps different analysis and problem-solving skills are required for the sections or the students were less familiar with the PT format of the assessment.

Table 1

Correlation Coefficients: PT and SRQ Scores. n = 6245 Students

	PT_APS	PT_WE	PT_WM	PT_TOT	SRQ_TOT
PT_APS	1.00				
PT_WE	.85**	1.00			
PT_WM	.66**	.74**	1.00		
PT_TOT	.92**	.95**	.87**	1.00	
SRQ_TOT	.28**	.27**	.23**	.29**	1.00

** Correlation is significant at the .01 level (2-tailed).

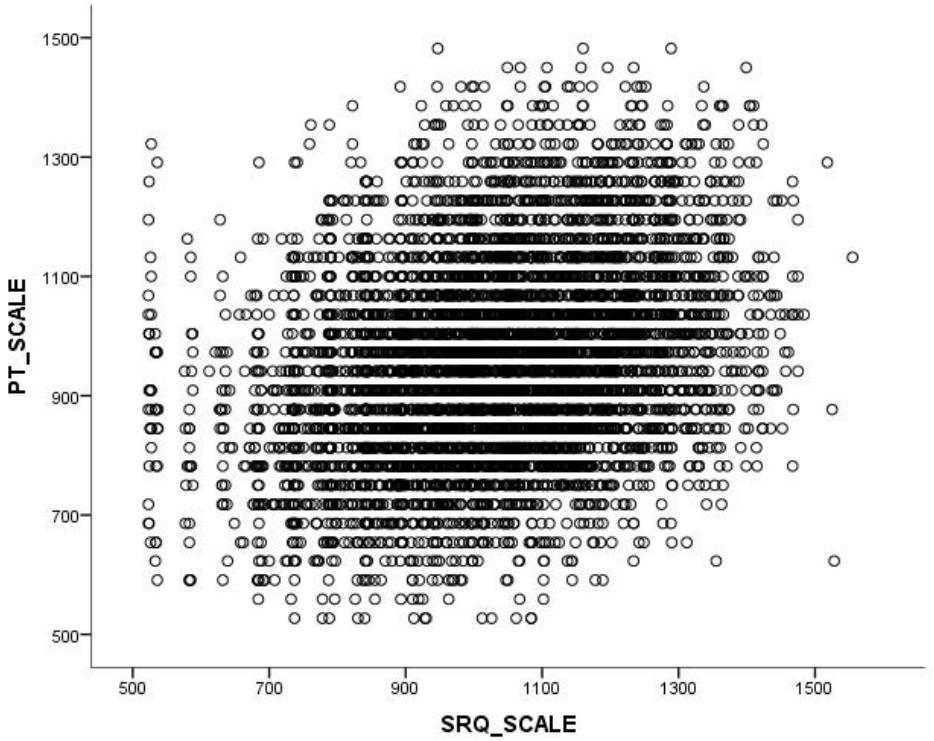

Figure 1. Correlation of SRQ and PT scale scores. n= 6245; r = .29.

At the institutional level, correlations between subsections are as expected (Table 2), indicating that institutions with students who have high PT scores also have students with high SRQ scores and vice versa.

Table 2
Correlation Coefficients: PT, SRQ, and SRQ Subsections at the Institutional Level. N = 23 Institutions

	PT_TOT	SRQ_TOT T	SQR	CRE	CA	TOT
PT_TOT	1.00					
SRQ_TOT	.85**	1.00				
SQR	.71**	.91**	1.00			
CRE	.85**	.86**	.60**	1.00		
CA	.72**	.91**	.79**	.75**	1.00	
TOT	.96**	.97**	.85**	.89**	.85**	1.00

** Correlation is significant at the .01 level (2-tailed).

The performances of the Italian and American students were compared (Table 3). Students performed comparably on the PT section of the assessment. The American students outperformed the Italian students by approximately half of a standard deviation on the SRQs (Table 4). Looking more closely at the subsections of the assessment, it appears that the American students outperformed the Italian students on the Critical Reading and Evaluation and Critique an Argument subsections, but not on the Scientific and Quantitative Reasoning subsection. The low correlation in scores between two sections of the assessment could be due to difference in performance.

Table 3
Descriptive Statistics for the CLA+ for Italian vs. American Students

		n	Mean	St. Dev.	Percentiles		
					25th	50th	75th
Total Scale Score	Italian Students	6245	995	135	905	993	1086
	American Students	11654	1116	150	1015	1121	1223
PT Scale Score	Italian Students	6245	958	164	845	941	1069
	American Students	11755	1093	168	976	1088	1207
SRQ Scale Score	Italian Students	6589	1033	172	916	1037	1152
	American Students	11959	1132	184	1004	1143	1269

Translation of student responses

CAE selected 25 Italian student PT responses that had perfect Italian scorer agreement. CAE then hired Capstan to back-translate and adapt these responses from Italian into English. The translations and adaptations were conducted to maintain the authenticity of the student responses. For example, if the student made a grammatical error in Italian, a similar error in English was made. The adaptation also included changing the cities back to their original names (Clinton and Greenville) rather than keeping Borgorosso and Borgoverde as the city names.

The 25 translated and adapted student responses were initially scored by two CLA+ scorers. The responses were mixed in with 25 American student responses to the same PT. The scorers were blind to the fact that half of the student responses were in fact back-translated Italian student responses. A third scorer was brought in to score 5 of the 25 responses because there was a difference of greater than two points between the initial two scorers. The average of the closest two scores was then used for the subsequent analyses. The inter-rater reliability as measured by the pearson correlation between the two total PT scores was $r = .97$, ; $p < .001$.

The correlation between the average total PT score for teams of American and Italian scorers was $r = .76$, $p < .01$. The Italian and American scorers had different mean scores for the 25 student responses (M.ITA = 9.72, SD.ITA = 5.13; M.USA = 11.06, SD.USA = 3.80). However, the average difference between the Italian scorers and the American scorers (M = 1.34; SD = 3.33) was not found to be significant ($t24 = 2.02$; $p = .055$; Figure 2. These results provides evidence that the scoring process, which includes scorer training, is valid and provides comparable results between the Italian and American teams.

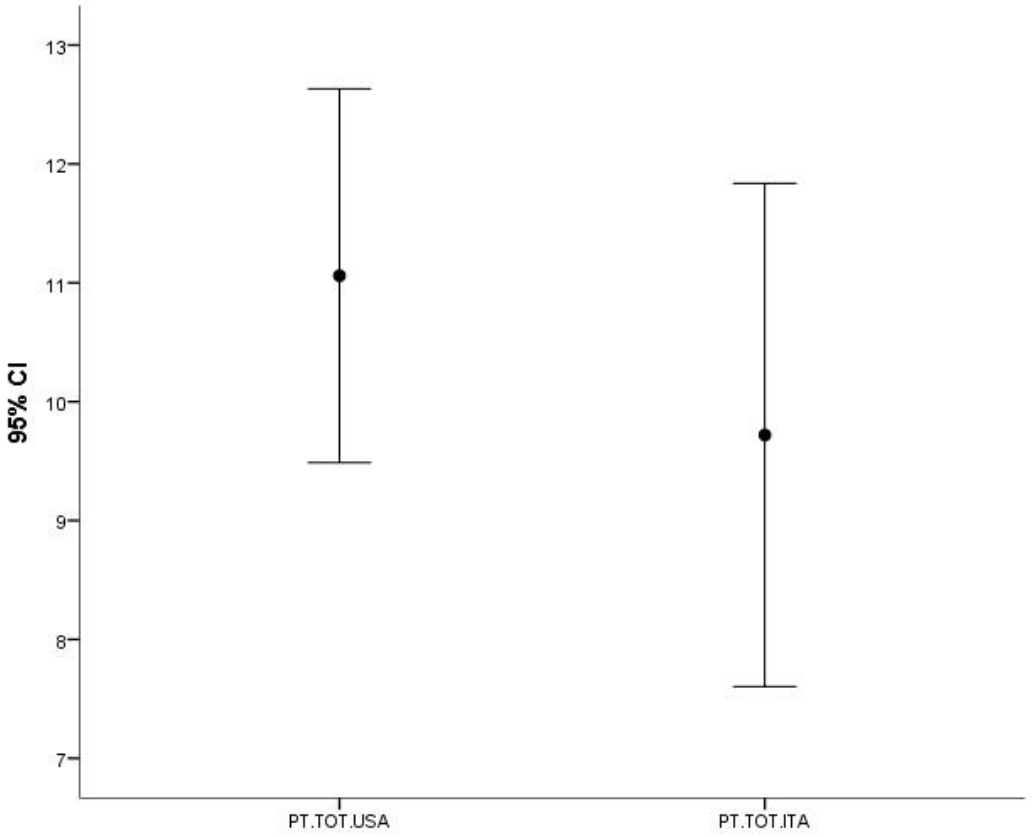

Figure 2. Mean PT total score with 95% CI error bars by scoring team, n = 25 student responses, t24 = 2.02; p = .055.

Discussion

When ANVUR implemented the second round of the CLA+, the purpose was to assess Italian students' critical-thinking and written-communication skills. The results show that the CLA+ can indeed be used to assess these skills and that the Italian students' performance on the CLA+ is roughly comparable to the results attained by their American counterparts.

One hypothesis is that Italian students are not familiar with taking standardized tests, let alone Performance Tasks, resulting in the correlation between the PT and SRQ scores being lower than expected. However, this is shown not to be true given the comparable PT scores between the two sets of students (Table 3). Regardless of the international comparison, there may be some merit to the hypothesis that the Italian students are not as familiar with standardized tests or PTs. One recommendation would be to develop a practice assessment for the students. This may help remediate issues pertaining to unfamiliarity with the format of the assessment.

At the institutional level, the PT, SRQ, and the SRQ subsections are highly correlated (Table 2), indicating that performance on the CLA+ is as expected when the data are aggregated to the institutional level. Institutions with students that have high PT scores also have students with high SRQ scores and vice versa. However, it should be noted that the SQR and CA sections had lower correlations than with all other sections, indicating that perhaps the quantitative and logical reasoning skills measured in these sections are different from the other critical-thinking skills assessed in the CLA+ for the Italian students.

Conclusion

Overall, results from this study indicate that the CLA+ measures the critical-thinking and written-communication skills of the Italian students. The reliability scores for each of the sections on the Italian version of CLA+ are comparable to the American version, and although overall reliability was low at the individual student level, this is not the case at the institutional level. The results from this study also indicate that Italian students' performance were comparable to that of their American counterparts.

The results from the 2015 administration of the Italian CLA+ corroborate results from the 2013

study, namely, that the CLA+ reliably measures the critical-thinking and written-communication skills of the Italian students.

Scientific significance

International assessments are challenging (Blömeke et al., 2013), but this study demonstrated that it is indeed feasible to measure critical-thinking and written-communication skills using a performance-based assessment. Reliability and validity can best be accomplished through a standardization process, and close collaboration is essential. Finally, employers in an international context will provide validity evidence that these higher-order skills are important for hiring decisions.

References

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C., & Fege, J. (2013). *Modeling and measuring competencies in higher education.* Springer.

Kahl, S. (2008). *The assessment of 21st century skills: Something old, something new, something borrowed.* Paper presented at the Council of Chief State School Officers 38th National Conference on Student Assessment, Orlando, FL.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: review and recommendations. *Educational Measurement: Issues and Practice*, 19, 5–15.