CAE

Evaluating Performance Task Scoring Comparability in an International Testing Program

March 27, 2014

Doris Zahner
Jeffrey T. Steedle

Abstract

The OECD launched AHELO in an effort to measure learning in international postsecondary education. This paper presents a study of scoring equivalence across nine countries for two translated and adapted performance tasks. Results reveal that scorers had similar notions of relative response quality, but not absolute quality.

Keywords: international, large-scale assessment, performance assessment, postsecondary assessment, scoring

Introduction

Performance based assessments are becoming increasingly more prominent in assessment programs (Kahl, 2008; Penfield & Lam, 2000). In the K-12 arena, the two large assessment consortia, Smarter Balanced and PARCC (PARCC, 2012; SBAC, 2012) have both incorporated performance-based items into their assessment plans. In fact, the creation of the Common Core State Standards confirms the nation's desire to shift the paradigm from content knowledge toward higher-order skills, such as integrating information across multiple sources, reasoning, and modeling. There is a parallel movement in higher education, particularly with online institutions, toward competency-based education (Burke, 2005). Performance tasks, which are open-ended assessments, require students to demonstrate their knowledge, skills, and abilities through constructing a response rather than answering a series of multiple-choice questions.

This shift toward performance-based assessment is not unique to the United States. When the Organisation for Economic Co-operation and Development (OECD) decided to conduct a feasibility study for the Assessment of Higher Education Learning Outcomes (AHELO), three different strands were chosen (AHELO, 2012a, 2012b). The first two, Engineering and Economics, are domain specific because they are tied to particular fields of study. The third, Generic Skills, is independent from any particular field of study because it focuses on the higher-order skills that are emphasized in current K-12 learning standards (e.g., the Common Core State Standards) and commonly listed as learning goals for postsecondary students in in the United States (Hart Research Associates, 2009).

The Collegiate Learning Assessment (CLA), the flagship assessment of CAE (Council for Aid to Education), was selected as the anchor for the Generic Skills strand of the AHELO feasibility study. This decision was based upon CLA's proven record of validly and reliably measuring generic (or general) skills in the United States. Additionally, the OECD found the analytic framework of the CLA to be a suitable starting place for the feasibility study. The OECD recognized that the core cognitive skills measured by the CLA (analytic reasoning and evaluation, problem solving, and written communication), using performance-based, open-ended prompts, accurately reflected the skills that the OECD wished to assess in the Generic Skills Strand of the AHELO Feasibility Study (Tremblay, Lalancette, & Roseveare, 2012).

The CLA Performance Tasks, which require students to solve simulated real-world problems using information provided in a document library, are designed to measure higher-order skills such as critical thinking and written communication. Such skills are recognized as essential for addressing the complex, non-routine challenges facing workers in the global Knowledge Economy (Hart Research Associates, 2006, 2009). In an effort to measure the attainment of these skills by postsecondary students, the OECD launched the AHELO Feasibility Study.

CAE's participation in the AHELO feasibility study consisted of overseeing the translation and adaptation of two performance tasks from the CLA and training international scorers on how to score the student responses. The performance tasks were adapted and translated from U.S. English for administration in eight other countries (Colombia, Egypt, Finland, Korea, Kuwait, Mexico, Norway, and the Slovak Republic). In order to establish cross-national norms using AHELO results, scores from different countries must be comparable. This paper presents the results of the analyses of the student responses from the nine participating countries (the regular administration) and a separate scoring equivalency study that was conducted using translated student responses for the two performance tasks.

Method

Regular Administration

The Generic Skills Strand of the AHELO project consisted of two sections: the Constructed Response Test (CRT) and the Multiple Choice Questions (MCQs). Two CLA performance tasks, "Lake to River" (LR) and "Catfish" (CA) were selected to be the tasks for the CRT section of the assessment. These tasks were chosen by a committee of representatives from the participating countries based on their adaptability for use in eight countries in addition to the United States. A detailed process for translating and adapting the tasks was implemented, including focus groups, field testing, and cognitive labs. Each CRT consisted of a scenario and a document library containing between 6 and 10 documents. Students were randomly assigned one of the two performance tasks. The students were expected to analyze the information presented in the documents, make inferences and connections, and draw a conclusion in order to address

the open-ended prompts. What is unique about CLA tasks is that there are multiple possible conclusions, solutions, or recommendations suggested in the document library. Each possible conclusion is supported by one or more pieces of evidence provided in the documents. Some of the possible conclusions may be better supported by the information in the documents than others. However, if a student chooses one of these less supported conclusions, he or she can still perform well on the task if his or her response is supported by information in the documents and written in a manner that conveys ideas clearly and logically. Students were given 90 minutes to read the documents and write (i.e., type) their responses to the prompts. A total of 10,657 student responses were scored in the feasibility study across the nine participating countries.

The student responses were scored using a modified version of the CLA scoring rubric. The CLA rubric had four subscores: Analytic Reasoning and Evaluation, Problem Solving, Writing Effectiveness, and Writing Mechanics. After much discussion and debate, the participating countries decided to eliminate Writing Mechanics (grammar, punctuation, vocabulary, and syntax) from the AHELO CLA scoring rubric due to cross-country differences for this scale.

The rubric subscores ranged from 1 to 6. Analytic Reasoning and Evaluation measured students' ability to interpret, analyze, and evaluate the quality of the information that was provided in the document library. This entailed identifying information that was relevant to a problem, highlighting connected and conflicted information, detecting flaws in logic and questionable assumptions, and explaining why information was credible, unreliable, or limited. The Problem Solving subscore assessed students' ability to consider and weigh information from discrete sources to make decisions, draw conclusions, or propose a course of action that logically follows from valid arguments, evidence, and examples. Students were also expected to consider the implications of their decisions and suggest additional research when appropriate. The Writing Effectiveness subscore evaluated how well students could construct an organized and logically cohesive argument by providing elaboration on facts or ideas. For example, students could explain how evidence bears on the problem by providing examples from the documents and emphasizing especially convincing evidence. Each student received a score from 1 – 6 on each of the subscores, so student total scores ranged from 3 – 18. If a student failed to respond to the task or gave a response that was off topic, a score of 0 (the equivalent of N/A) was given for all three subscales, and these student responses were eliminated from any subsequent analyses.

The CRTs were administered online via a secure testing platform during an assigned testing window ranging from late 2011 into early 2012. Each country then selected one or two lead scorers to attend in-person trainings on how to score the student responses to the performance tasks. The main goal of the scorer training was to calibrate the scorers and establish consistent scores across the countries for the same student response. The lead scorers then conducted scorer trainings within their own countries. Following the full test administration, each student response was scored twice by trained scorers in each country.

Scoring Equivalency Study

In addition to scoring the student responses from each individual country, a set of scoring equivalency papers were translated and then randomly distributed into the score queues of participating countries. Only five of the eight countries, including the US, participated in the scoring equivalency study. The scoring equivalency study involved two samples of student responses: Sample A and Sample B. Sample A (N=209) consisted of up to 100 papers initially written in English and translated into four additional languages. The scorers from the four participating countries were unaware that there were translated student responses in their scoring queues. Sample B (N=121) consisted of up to 51 papers from the participating countries initially written in students' native languages and translated into English. These student responses were then inserted into the American scorers' queue. Like the scorers from the participating countries for Sample A, none of the scorers from the US knew that there were translated papers in the scoring queue.

Scorer agreement within countries was examined by calculating correlations between scorers. The scoring equivalence data were analyzed by comparing mean scores on common sets of translated responses across countries.

Results

Regular Administration

First, how reliable was scoring within countries? The students' native language responses from all nine countries (N = 10,657) were used in the analyses. The within-country agreement (i.e., correlations between two scorers) ranged from .52 to .88 (median .77). Correlations between two pairs of scores ranged from .68 to .94 (median .87).

Figure 1 reveals that, with one notable exception, the relative standings of the countries were similar on the two PTs, but the mean LR and CA scores were different for most countries. This could be a reflection of a difference in task difficulty or perhaps scorer leniency. If the CRTs were equally difficult and the scorers scored with equal severity within each country, the mean LR and CA scores would be more closely aligned to the diagonal.

In nearly all countries, the mean LR score was higher than the mean CA score, suggesting that LR was an easier task than CA. However, some component of the differences between mean LR and CA scores is attributable to differences in scorer severity between the tasks. Consider, for example, the country represented by the color green. In this country, the mean LR score was much higher than the mean CA score. Some of this was attributable to the difference in task difficulty, but the LR scorers in this country were likely more lenient than the CA scorers, which would have inflated the difference between the LR and CA means.

To adjust for these differences, equipercentile equating with a random groups design was employed. This equating design was made possible by the fact that CRTs were randomly assigned to the participating students. The equipercentile method was employed to equate CA to LR in each country separately. Mean LR and equated CA scores are shown in the right panel of Figure 1. As expected, the mean CRT scores were much more consistent after equating. The relative standings of countries shown in the right panel of Figure 1, however, do not necessarily reflect the "true" relative standings of countries on the CRTs since the relative standings shown in Figure 1 are highly dependent on which CRT was chosen as the "base form" (LR was arbitrarily chosen as the base form in this analysis).

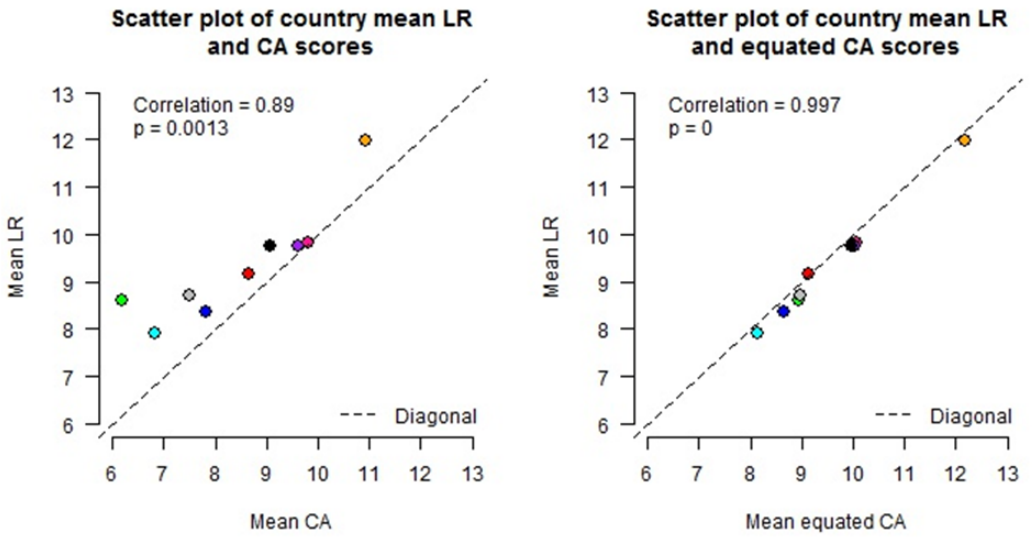

Figure 1. Scatter plot of each country's mean LR and CA scores before and after equating the CRTs in each country.

As previously stated, there were two different sections to the Generic Skills strand of the AHELO feasibility study, the CRT section and the MCQ section. The correlations between CRT and MCQ scores were computed and the range of correlations observed is shown in Table 1. The correlation between LR and MCQ scores ranged from .16 to .61. For CA, the correlation ranged from .23 to .55. However, the correlations obtained when combining all data were somewhat higher (.45 and .47). When data from both CRTs and all countries were combined, the correlation between CRT and MCQ was .45. A graphical representation of this linear association is provided in Figure 2.

Table 1: *Correlations between CRT total scores and MCQ by country*

Country	LR	CA	CRT
A	.40	.40	.40
B	.48	.40	.41
C	.26	.23	.24
D	.28	.29	.28
E	.36	.28	.32
F	.61	.30	.45
G	.38	.47	.42
H	.16	.33	.24
I	.57	.55	.56
Total	**.45**	**.47**	**.45**

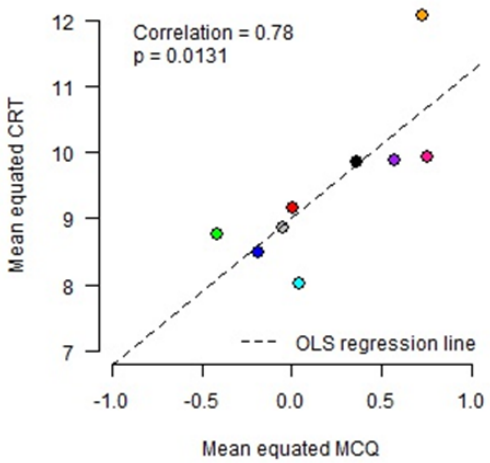

Figure 2. Mean PT (CRT) versus MCQ scores.

When comparing mean CRT to MCQ scores (Figure 2), three countries (the green, cyan, and orange points in Figure 2) showed large differences between observed and expected PT means. Otherwise, the relative standings of each country's mean CRT scores appeared as expected given their mean MCQ scores. Some countries had mean CRT scores that were well below or well above what would have been expected given their mean MCQ scores, suggesting again that there were between-country differences in the judgment of the absolute quality of the CRT responses. These differences could be a reflection of measurement error in the CRT and MCQ scores. Another possible explanation is that the CRTs and MCQs are not measuring the same construct or abilities. If this is true, then there should not be a perfect linear relationship between the CRT and MCQ scores. Differences in scorer severity or leniency is also a possible explanation for the difference in the observed versus expected student performance for the two sections. For example, scorers in the country represented by the yellow point may have been particularly lenient. Other between-country differences in the testing conditions (e.g., difference due to translation and adaptation, differences in student motivation, recruitment, sampling, incentives, and test administration conditions) may have also influenced student results. Thus, the relative standings of countries shown in Figure 2 do not necessarily reflect their "true" relative standings. From these results, it should be observed that there exist unexplained differences between observed CRT performance and expected CRT performance (based on average MCQ performance). Differences in scoring is a plausible explanation for such differences.

Scoring Equivalency Study

In addition to investigating whether CRT results could be reliably scored in a standardized international testing environment, analyses were conducted to investigate whether student responses received the same relative and absolute scores regardless of language or country. "Sameness" was examined in two ways: relative and absolute. The first refers to whether the relative standings of the responses were consistent (i.e., highly correlated) regardless of language or country. The second reflects whether the mean scores of the responses were equal.

Two sets of sample student responses were created. Sample A (N=209) consisted of up to 100 papers initially written in English and translated into four other languages. Sample B (N=121) consisted of up to 51 papers from the participating countries initially written in students' native languages and then translated into English. Both sets of student responses were inserted in the appropriate queues for scoring.

Relative Quality of Scores

To determine whether the scorers agreed with one another about the relative quality of the responses, correlation coefficients among pairs of scorers, both within and across countries, were calculated. The within-country analyses were presented earlier in this paper and showed that student responses could be reliably scored within each country.

For the Scoring Equivalency study, each student response was scored by a total of four scorers, two from the United States and two from another country. As shown in Table 2, the inter-country correlations of the total scores, for two scorers, in each country are relatively high, with r = .65 to .92 for Sample A and r = .57 to .96 for Sample B, for both performance tasks across all countries. The exception is Country 4 for Sample B. Based on these correlations, if the results from Country 4 for Sample B are removed, it appears that the relative quality of a response, regardless of the language in which it was initially or subsequently scored, can be reliably determined. Further investigation into the low correlation between Country 4's scorers for the Sample B responses is needed.

Table 2: Inter-country scorer agreement

	Catfish		Lake to River	
Country	**Sample A**	**Sample B**	**Sample A**	**Sample B**
Country 1	.78	.96	.92	.94
Country 2	.88	.91	.88	.95
Country 3	.65	.70	.78	.84
Country 4	.88	.95	.78	.57
Mean	**.80**	**.88**	**.84**	**.83**

Absolute Quality of Scores

To determine whether the scorers agreed with one another about the absolute quality of the scores, the mean scores across countries were analyzed. Figure 3 illustrates the mean scores for each country on the same 30 open-ended responses from Sample A. Results corroborate the observed differences in the difficulty between the two CRTs from the regular administration of the assessment (Figure 1).

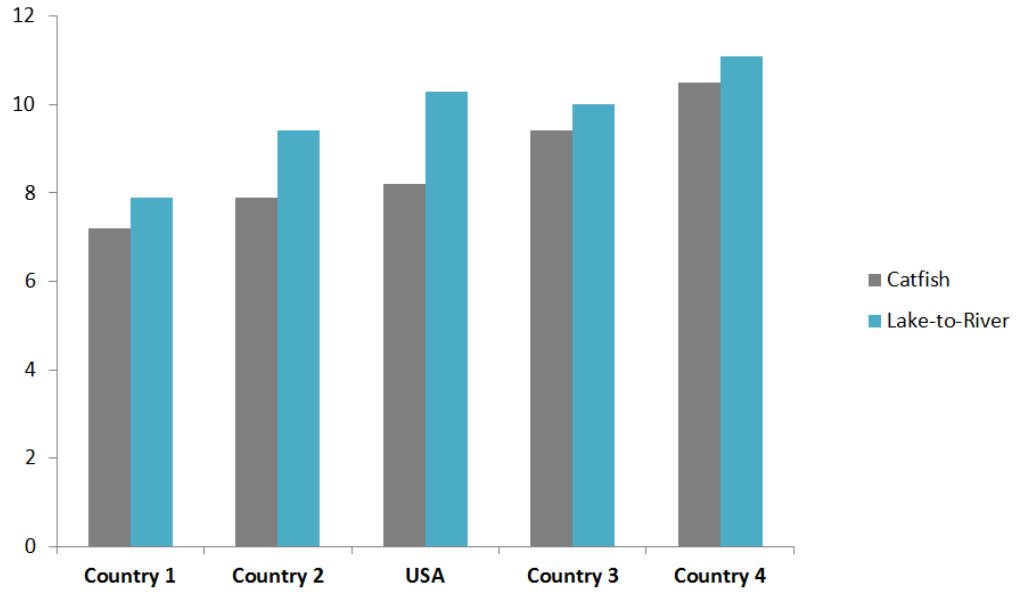

Figure 3: Mean scores for each country on the same 30 open-ended responses (Sample A)

However, there appears to be a difference in the mean scores for the 30 papers in Sample A across the countries, indicating that there were systematic differences across countries in terms of absolute quality ratings, $F_{4,112} = 24.32$; $p < .0001$ (recall that these are the same student responses scored in different countries). The pattern of mean scores suggested that translating responses affected the perceived response quality or that there was a difference in scorer leniency across countries.

Moreover, results suggest that translated responses are not scored the same as responses originally composed in the native language of the scorer. For example, scorers from Country 3 rated student responses from Sample A (papers originally written in English and translated into different languages) an average of 1.4 points higher than those given by the U.S. scorers. One would expect the same average difference between these scorers for student responses from Sample B, where the student responses were originally written in their native language and then translated into English, but this did not occur. Instead these same scorers from Country 3 rated the papers from Sample B an average of 2.2 points higher than the American scorers. This pattern of results, where the average difference in scores for papers from Sample B is greater than the average difference in scores for papers from Sample A, arose for all countries on both performance tasks. This pattern is consistent with the notion that scorers are less favorable toward translated responses, more favorable toward untranslated responses, or both. This finding raises concerns about possible future use of this method for comparing the absolute quality of performance task responses across languages.

Conclusion

This study of scoring equivalency across languages and countries provides several findings of interest to international assessment programs. The first is that scoring reliability within countries was high, indicating that scorer training was effective within each country participating in the AHELO feasibility study. Thus, students can be assessed on their higher-order skills using an open-ended assessment within a given country. However, when comparing results across countries, there were notable between-country differences in the judgment of the absolute quality of the student responses to the CRTs.

When looking specifically at student responses that were translated from English into other languages (Sample A) or from other languages into English (Sample B), a similar pattern was found. The scores assigned to responses varied across the countries but were highly consistent within a country (Table 2). Additionally, there appears to be a high degree of agreement within, and between, countries on task difficulty (LR is easier than CA) and with the relative answer quality. That is, although the scores that were assigned to the student responses varied across countries, the scorers were very consistent in the rank ordering of the quality of the responses across countries.

The results indicate that there are differences in scorer leniency across the countries on the open-ended assessment. This is not an unusual result, and there are ways to account for the difference in mean scores. One way is to equate the open-ended results using the results from a more objective assessment such as a multiple-choice test.

A final finding from this study is that it is feasible to develop, translate, administer, and score the responses to a computer-based, college-level, open-ended assessment of general knowledge, skills, and abilities that are applicable to many countries. Scores from an international testing program can be calibrated to recognize relative response quality, but not absolute response quality. The scores on such tests can provide valid and reliable data for large-scale international studies. CAE does not recommend that these measures be used to make judgments about individual students, but they can be used in large-scale assessment programs, if one is interested in how well students are performing on average. With the increasing popularity of performance-based assessments and a global interest in critical-thinking skills, it makes sense to further develop an international assessment such as the AHELO test of Generic Skills.

In the future, we hope to complete the study by examining the in-depth relationship between the scores on the multiple-choice and open-ended assessments. Additionally, we hope to conduct studies to support a large-scale international implementation of an open-ended assessment measuring critical-thinking and written-communication skills.

References

AHELO. (2012a). AHELO feasibility study interim report. Paris: OECD.

AHELO. (2012b). Testing students and university performance globally: OECD's AHELO. Retrieved February 28,2014, 2014, from http://www.oecd.org/edu/skills-beyond-school/testingstudentanduniversityperformancegloballyoecdsahelo.htm

Burke, J. (Ed.). (2005). Competency based education and training: Routledge.

Hart Research Associates. (2006). How Should Colleges Prepare Students to Succeed in Today's Global Economy? - Based on Surveys Among Employers and Recent College Graduates. Washington, DC: Hart Research Associates.

Hart Research Associates. (2009). Learning and Assessment: Trends in Undergraduate Education - A Survey Among Members of The Association of American Colleges and Universities. Washington, DC: Hart Research Associates.

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.

PARCC. (2012). Partnership for Assessment of Readiness for College and Careers. Retrieved November 20, 2012, 2012, from http://www.parcconline.org/about-parcc

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.

SBAC. (2012). Smarter Balanced Assessment Consortium. Retrieved February 28, 2014, 2012, from http://www.smarterbalanced.org/

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). Assessment of higher education learning outcomers (AHELO) feasibility study report: Design and implementation (Vol. 1).