CAE

Methodological Challenges in International Comparative Post-Secondary Assessment Programs: Lessons Learned and the Road Ahead

Raffaela Wolf
Doris Zahner
Roger Benjamin

Wolf, R., Zahner, D., & Benjamin, R. (2015). Methodological challenges in international comparative post-secondary assessment programs: lessons learned and the road ahead. *Studies in Higher Education*, *40*(3), 471-481.

Introduction

There is substantial merit in evaluating learning outcomes in higher education in the twenty-first century. As shifts in skill sets and human capital are gaining in importance for economic prosperity, growing numbers of students enroll in post-secondary education programs with the hopes of preparing adequately to meet the challenges of constantly changing economies and societies. Students' learning outcomes can and should be utilized as a key factor when evaluating institutional, as well as individual, performance. While information on an institution's contribution to students' learning may be obtained indirectly through graduate- and student-engagement surveys, to date, the use of measurement instruments as a means to assess learning outcomes globally is still in its infancy.

Since its inception in 1959, the International Association for the Evaluation of Educational Achievement (IEA) has conducted a series of international comparative studies with the aim to provide policy makers, educators, and researchers with insight regarding the educational achievement of students and learning contexts. However, the majority of these studies focus on assessing student achievement in the secondary-school sector. A new movement towards an innovative international research initiative in higher education, the Assessment of Higher Education Learning Outcomes (AHELO) feasibility study, has been sponsored by the Organisation for Economic Co-operation and Development (OECD, 2011). The main premise of the AHELO study is to develop a cross-national concept for valid assessment of generic and domain-specific learning outcomes of students from diverse degree courses, programs of study, and higher-education systems on an internationally comparative basis (Tremblay et al. 2012; 2013). Through the AHELO feasibility study (OECD, 2012, 2013), it is evident that the purpose of the study is multi-faceted. Consequently, numerous conceptual and methodological challenges need to be addressed when employing competency-based assessments within cross-cultural contexts.

The AHELO feasibility study included three strands (AHELO, 2012b), two domain-specific strands (engineering and economics) and one Generic Skills Strand. The Generic Skills Strand is independent from any particular field of study because it focuses on higher-order skills that are typically regarded as important learning objectives for post-secondary students in the United States (Hart Research Associates, 2009). Three U.S.-based testing organizations have designed assessments that target higher-order skills for over a decade. These tests are the Collegiate Learning Assessment+ (CLA+; CAE), the ETS Proficiency Profile (EPP; ETS), and the Collegiate Assessment of Academic Proficiency (CAAP Program Management).

The Collegiate Learning Assessment (the original manifestation of CLA+), the flagship assessment of CAE, was selected as the anchor for the Generic Skills Strand of the AHELO feasibility study. The CLA is an open-ended, performance-based measure that purports to measure higher-order skills, such as analytic reasoning and evaluation, problem solving, and written communication. OECD launched the AHELO feasibility study in an attempt to measure whether these types of skills are achieved by students in the tertiary education sector. The Generic Skills Strand of the AHELO project consisted of two sections: the Constructed Response Test (CRT) and the Multiple Choice Questions (MCQs). Both sections were used under the assumption that a similar construct was measured regardless of item format.

AHELO and other international comparative assessment systems face numerous methodological challenges due to heterogeneity in educational systems, socio-economic factors, and perceptions as to which learning outcomes should be assessed within a higher-education framework. These aforementioned complexities pose psychometric challenges that pertain to test design and development, translation, adaptation, student sampling, scoring, reporting, and the validity of score interpretations.

The goal of this paper is to generate ideas for the improvement of cross-national research agendas, such as the AHELO project. The main purpose is to focus on the lessons learned from the AHELO feasibility and other international studies that would help inform the research of future multi-national educational assessment studies. First, the observations and findings of six specific methodological points from the AHELO study's Generic Skills Strand design are reiterated and discussed (Klein, Zahner, Benjamin, Bolus, & Steedle, 2013). Next, the findings of two international case studies are outlined and discussed. Lastly, general methodological considerations in cross-national, higher-education research designs are considered and reviewed.

Methodological Constraints Discovered with the AHELO Feasibility Study

In theory, the goal of sampling for an international assessment is to test a representative sample of students at all ability levels across countries. However, in reality, there are, inevitably, major practical sampling problems despite the best intentions and most sophisticated sampling designs.

Consequently it is unlikely that comparable samples of students, schools, and regions are tested across countries (Education Week, 2008). The Generic Skills Strand of the AHELO feasibility study planned to use a random sample of students in every participating country. However, AHELO was not able to test a national probability sample in any country because many schools and students refused to participate in the no-stakes test, and there is no satisfactory analytic way to adjust for that problem. Consequently, AHELO could not provide countries (or schools) with meaningful benchmarks or norms against which they could compare their students' performance to that of students in other countries or institutions.

The sampling problem could have been mitigated by identifying and gathering data about several variables that are likely to influence scores in all countries and then using a case-mix control and weighting to adjust for differences between the samples' characteristics and those in the corresponding populations from which they were drawn. Prior research has established that motivation is significantly related to test performance (Cole, Bergin, & Whittaker, 2008; Wise & DeMars, 2005), and some experimental evidence indicates that students are less motivated and perform worse in low-stakes testing conditions (Cole & Osterlind, 2008).

Additionally, the AHELO feasibility study showed that, on average, students who said they tried to do their best on the assessment scored significantly higher than those who did not indicate as much effort. Since a matrix sampling approach with no stakes attached to higher performance was employed in the feasibility study (AHELO, 2012a), differences among countries and schools in the effectiveness of their educational programs may be confounded by differences between their students' motivation, ability, and willingness to participate and complete the assessment. However, other factors (e.g., item difficulty, test time, and test fatigue), either alone or in combination, may explain why many students did not finish the assessment, despite no penalty for guessing or writing a minimal amount of text to answer the constructed-response tasks. Furthermore, those involved in developing solutions to the motivation issue may want to employ a protocol that will involve some stakes for the participants. Also, the unit of analysis is important for the motivation problem as a student-level analysis with stakes is essential to motivate students to do their best.

The AHELO project recognized that the validity of its scores rested on a very strong assumption. Specifically, it was assumed that these scores could be interpreted the same way across countries regardless of the language in which the questions and their answers were written. AHELO assumed that an item's difficulty and the construct it measured was not affected by the language in which the item was written or by the translation process. This assumption is crucial to establishing validity. Based on Differential Item Functioning (DIF) analyses (Klein et al., 2013), it appeared that this assumption was violated for most countries that took Australian's Council of Educational Research (ACER's) Generic Skills Strand multiple-choice tests and, in most countries, it was violated for nearly half the items. If the assumption that scores may be interpreted the same way across countries cannot be corroborated, a different set of goals, methodologies, and research strategies needs to be developed.

Another challenge to international testing of constructed-response tests, as exhibited in the AHELO feasibility study, is that open-ended tasks may be influenced by differences in score leniency. Scorers, either within a country or across countries, may not score the same student response consistently or equivalently. However, score equivalency across countries is essential if scores are to be compared across countries. CAE's previous scoring-equivalency analyses (Klein et al., 2013; Zahner & Steedle, 2014) indicated that scorers in different countries had similar notions of the relative quality of CRT responses, but there were between-country differences in notions of the absolute quality of CRT responses (i.e., differences in scorer severity). However, this does not address the sampling, participation, and motivation issues noted above.

The reporting of subscores can be useful at the individual student level as well as the institutional level. Subscores at the individual student level may aid students in identifying their own strength and weaknesses, which may be used to guide future remedial work. At the institutional level, subscores may be utilized to create a profile of performance for their graduates, which may aid in evaluating institutional effectiveness. This information can be used to identify areas that are necessary for instructional improvement. AHELO's Generic Skills Strand's multiple-choice test initially had 55 items, but only 52 of them were administered. The administered items were randomly divided into four subtests, with each subtest having 23 or 24 items. There were 14 items common to all four subtests. This division and the descriptions of the subtests suggest they assess somewhat different constructs, but no empirical evidence is provided to support that conclusion. At a minimum, the AHELO report should provide the correlations among the item sets before and after being adjusted for their reliabilities (i.e., corrected for attenuation). Subscores are more likely to add value when they have a high reliability and the correlation between the true subscore and the true total score is low (S. J. Haberman, 2008; S. Haberman, Sinharay, & Puhan, 2009; Puhan, Sinharay, Haberman, & Larkin, 2010). Furthermore, a moderate correlation between the true subscore and true total score may be an indication that the subscore is contributing unique information over and above the total score alone, hence adding value to the overall test-performance interpretation.

Assessments that are composed of multiple item formats (i.e. performance tasks and selected-response items) may assess different constructs despite the intention of measuring an unidimensional latent ability. For example, ACER's analytic-reasoning test assesses different abilities than CAE's constructed-response tests even though both tests are billed as measuring the "same" construct, and there is only a moderate correlation between them even when the school is the unit of analysis. This may have occurred as a result of a student's ability in one area being correlated with that student's ability in another area.

CAE conducted a test-validity study that examined whether commonly used measures of college-level general educational outcomes provided comparable information about student learning (Klein et al., 2009). Specifically, did the students and schools earning high scores on one such test also tend to earn high scores on other tests designed to assess the same or different skills? And, were the strengths of these relationships related to the particular tests used, the skills (or "constructs") these tests are designed to measure (e.g., critical thinking, mathematics, or writing), the format they use to assess these skills (multiple-choice or constructed-response), or the tests' publishers? Results from the test validity study indicated that the unit of analysis had a much greater impact on correlations than did response mode or construct. Furthermore, score reliability is much higher when the school—rather than the student—is the unit of analysis (Klein et al., 2009). This is important because the stated goal of AHELO is to provide reliable and valid assessments, cross-nationally, that will assist in the improvement of teaching and learning of institutions.

Summary and Implications

The observations based on the analyses of the AHELO Generic Skills Strand findings (Klein et al., 2013; Zahner & Steedle, 2014) presented in this paper suggest that the equivalency assumption of comparable results is difficult to attain in cross-national research endeavors, regardless of item type or format. Thus, caution needs to be taken when score inferences are to be drawn across countries. While there may be differences in scorer leniency across the countries on the open-ended tasks, it is possible to account for the difference in mean scores through equating methodologies. The type of test that is given to students matters and so does the level of analysis. Although high correlations between the MCQs and the CRTs at the institutional level, as opposed to the student level, were observed in some studies, this does not imply that the two types of assessments are measuring the same constructs. Rather, students who perform well on one type of task tend to perform well on other types of tasks as well. The goal of attaining a national probability sample was not achieved. Recruitment and motivation problems limited the confidence one can have in the findings. Stakes will be needed to attract and motivate students to take the test and do well on it. Each of these issues is significant enough to warrant sustained research on its own. It will take considerable time and effort to develop research strategies that can assist in generating practical results that can be seen to be both reliable and valid.

Two recent case studies examined further issues within the context of international assessment programs (Wolf, Zahner, Kostoris, & Benjamin, 2014; Zahner & Steedle, 2014). Zahner and Steedle (2014) examined scoring equivalence across nine countries for two translated and adapted performance tasks. 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. Zahner & Steedle investigated whether open-ended response results could be reliably scored in a standardized international testing environment and whether student responses received the same scores regardless of language or country. Results indicated that scoring reliability within countries was high, indicating that scorer training was effective within each country participating in the AHELO feasibility study. 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 open-ended responses.

In late 2012, the Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR) approached CAE proposing a research study to test the feasibility of adapting, translating, and administering CLA+ to higher-education students in Italy. The purpose of this feasibility study was twofold. The first purpose was to see if it was possible to assess Italian students' higher-order skills as outlined in Table 1.

Task	Subscales
CRT	Analysis and Problem Solving
	Writing Effectiveness
	Writing Mechanics
MCQ	Critical Reading and Evaluation
	Scientific and Quantitative Reasoning
	Critique an Argument

Table 1. CLA+ Tasks and Subscales

The second purpose was to see if the Italian students' performance was comparable to their American counterparts. More specifically, Wolf, Zahner, Kostoris, and Benjamin (2014) examined whether students from Italy and the US ascribe the same meanings to different item formats (performance tasks and SRQs) thus addressing the issue of measurement equivalence (at the scale level) and the feasibility of cross-national score comparisons.

Results indicated that Italian and American students appeared to associate the same meaning to the definition of higher-order skills and that the items on the instrument were adequately sampled from the domain of higher-order skills. This finding ensured construct representativeness and the feasibility of measuring higher-order skills across countries. However, students appeared to associate different meanings with different item types across countries, which imposes the question as to whether valid score inference can be drawn from direct score comparisons of students in different countries.

Both studies illustrate that it is imperative to assess whether the cross-country equivalency has been met. However, in practice it may be difficult to attain the equivalency across countries. Consequently, the complexities of the skills being measured may not be truly unidimensional, which urges us to consider further avenues of research design within cross-national comparative studies.

Methodological Considerations in Cross-national, Higher-Education Research Design

AHELO, amongst other international assessment programs, rests upon the assumption that education is essentially a pragmatic venture with the fundamental goal of educational research to inform stakeholders in regards to pedagogy, curricula, practices, and schooling systems (Raudenbush, 2005). In order to fulfill this purpose, educational researchers rely on designs that allow inferences in regards to causal mechanisms. For example, a researcher may be interested in examining "what" causes differences in test performance across countries. An important feature of these types of designs is the control for unobserved heterogeneity in the sample. More specifically, the identification of causal effects may be corrupted because individuals that have attained equivalent degrees of education may show big discrepancies or heterogeneity in skill sets, both within and between countries. It becomes evident that theoretical models must be established to account for this within- and between-country heterogeneity. The construction of sound psychometric models and the inclusion of theoretically sound background variables are the building blocks in this endeavor. The measurement of those hypothetical constructs provides data that can be utilized to inform about educational effectiveness and decision making.

As established by the AHELO feasibility study, pure experimental design studies are often not feasible due to the nature of the setting, thereby limiting the extent to which valid causal inferences can be drawn. Over the past two decades new statistical techniques have emerged within the field of educational research in order to address the aforementioned complexity. These procedures include propensity score matching (Rosenbaum & Rubin, 1983, 1984) and instrumental variables assuming observational data with the premise of a quasi-experimental design. Within an educational research context, the collection of relevant background variables that are related to student achievement are often integrated in the research design. Students may differ substantially on these covariates and, thus, it may be useful to balance the distributions of the observed covariates between the treatment and comparison groups prior to conducting statistical analyses of treatment effects. For a relevant set of covariates, a scalar score (i.e. propensity score) can be created that summarizes the information obtained from the set of covariates. The groups involved in the analysis can then be balanced on this scalar or propensity score.

Educational researchers have utilized the application of propensity score matching in numerous ways, such as matching, stratification, and regression adjustment. Detailed examples within an educational context are given by several authors (Brand & Halaby, 2006; Byun, 2010; Cook, Steiner, & Pohl, 2009; Frisco, Muller, & Frank, 2007; Hong & Yu, 2008; Ou & Reynolds, 2010). Another strategy includes the application of instrumental variables (Angrist & Krueger, 2001). A successful implementation of these methods is contingent upon the quality of the background variables. The implementation of these methods is ad-hoc in nature and, thus, in order to identify causal relations, it is imperative to include instrumental variables into the assessment framework a priori. The advances in methodologies for making causal inferences from observational data suggest that there are procedures that facilitate the examination of substantive research questions within the field of education. However, it is evident that each method by itself has limitations, which implies that the use of multiple approaches may be more appropriate in order to ensure valid causal inferences. Furthermore, new multidisciplinary approaches to analyzing complex data from international comparative studies may provide new insights and solutions to complex problems outlined in this paper. Traditionally, economists, sociologists, and political scientists address research questions at the higher levels of the educational system. Conversely, educational scientists and psychologists are typically focused on individual differences between students or teachers; thus, these analyses are conducted at the lower level. Consequently, the different disciplines employ different methodologies to address research questions.

The AHELO project was initiated to accomplish numerous goals. Individuals involved with the design of the AHELO project recognized that in order to ensure valid and reliable assessment results, each instrument must match the objectives of the assessment. Furthermore, the construction of a valid and reliable measurement instrument is threatened by numerous sources of disturbances, such as systematic and random errors of measurement. A recommendation to mitigate the different sources of bias within the assessment design phase would be to devote efforts focusing on smaller, purposeful, and targeted assessment programs that aim at attaining a narrow but well-defined set of objectives rather than focusing on creating a large-scale assessment system that aims to solve all problems at once but does not yield interpretable results. For example, one may want to focus on examining the growth in effect size of student learning gains at institutions within countries. The changes in the amount of growth can then be compared across countries. Post-secondary academic institutions differ in their admissions criteria and requirements; consequently, part of the differences on observed outcomes may simply reflect the differences in input. For the purpose of examining value added at the institution level, it is imperative to implement a longitudinal design with repeated measures of competence at the commencement and completion of higher education. This type of analysis would focus on the institution as the unit of analysis.

Based on the observations from the AHELO feasibility study, it appears that numerous latent constructs are being measured. Employing statistical methodologies that rest upon the assumption of unidimensionality (i.e. a single construct is being measured) to score the assessment would likely yield inaccurate results and, thus, limit the generalizability and validity of any score inferences. Several recent developments within an IRT framework hold promise for assessing complex competencies. These advancements include: explanatory IRT models, multidimensional IRT models, and cognitive diagnostic models (Hartig & Hoehler, 2008). Additionally, the development of structural equation modeling (SEM) encompasses the use of a large number of statistical models to evaluate the validity of theoretical conceptions with empirical data (Hoyle, 1995). Within the context of international comparative studies, SEM has been an effective resource in addressing the nested (hierarchical) structure of units of educational systems, where examinees are nested within institutions and institutions are nested within countries.

As noted earlier, the unit of analysis impacts the reliability of scores. The majority of literature on cross-country research has focused on examining psychometric properties of instruments at the individual student level. A conundrum is that data for cross-country comparisons fundamentally follow a hierarchical structure in that individuals are nested within institutions and institutions are nested within countries; therefore, in order to ensure sound psychometric practices, psychometric properties at the country level should not be neglected. Statistical procedures that address both the individual (i.e., disaggregated) and the group (i.e., aggregated) levels do exist but have rarely been applied within a cross-country comparison context. Consequently, single-level data structures are commonly used to make inferences regarding country-level comparisons. However, caution is warranted because failure to account for dependencies in the data may limit the validity of the inferences to be drawn and may lead to distorted results. Growth over time has also been examined through the SEM framework. Furthermore, multiple sample-specific models can be examined simultaneously and sampling weights that account for the sampling design can be included in the analysis. SEM can be utilized with experimental and non-experimental data as well as cross-sectional and longitudinal research designs.

Regarding experimentation with twenty-first century assessments, the largest R&D program for development of new assessments, based on $360 million from the U.S. Department of Education, is in its second major phase of implementation. U.S. testing organizations, such as ETS, Pearson, CBT-McGraw-Hill, and CAE, are developing thousands of performance tasks, selected–response items, and other tests that take advantage of education technology. Consequently, different test types and testing protocols can be developed to improve teaching and learning in colleges and universities. Countries participating in international comparative studies may benefit from this opportunity by learning how to develop their own measurement instruments in collaboration with measurement scientists from the testing organizations as noted above.

Lastly, it may also be useful to approach a different research topic, such as work readiness. Political and economic leaders are focused on improving their human capital primarily because they want their work forces to assist the private sector in their countries in competing globally. The question to focus on, therefore, is what skills do employers find to be most important? One might start by undertaking studies of this question among similarly situated countries, such as those with the most advanced economies. The companies become the unit of analysis, which obviates the need for national probability samples and cross-national equivalency assumptions. A cross-national working group of experts might oversee the assessments to be developed to measure work readiness and tests administered in assessment centers in each participating countries. This type of research approach may be extended by focusing on establishing a connection between graduating seniors and employers (Benjamin, 2014).

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

Angrist, Joshua, & Krueger, Alan B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments: National Bureau of Economic Research.

Benjamin, Roger. (2014). Leveling the Playing Field From College to Career: Council for Aid to Education.

Brand, Jennie E, & Halaby, Charles N. (2006). Regression and matching estimates of the effects of elite college attendance on educational and career achievement. *Social Science Research, 35*(3), 749-770.

Byun, Soo-yong. (2010). Does policy matter in shadow education spending? Revisiting the effects of the high school equalization policy in South Korea. *Asia Pacific Education Review, 11*(1), 83-96.

Cole, James S., Bergin, David A., & Whittaker, Tiffany A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*, 609-624.

Cole, James S., & Osterlind, Steven J. (2008). Investigating Differences Between Low- and High-Stakes Test Performance on a General Education Exam. *The Journal of General Education, 57*(2), 119-130.

Cook, Thomas D, Steiner, Peter M, & Pohl, Steffi. (2009). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research, 44*(6), 828-847.

Frisco, Michelle L, Muller, Chandra, & Frank, Kenneth. (2007). Parents' union dissolution and adolescents' school performance: Comparing methodological approaches. *Journal of Marriage and Family, 69*(3), 721-741.

Haberman, Shelby J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204-229.

Haberman, Shelby, Sinharay, Sandip, & Puhan, Gautam. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*(1), 79-95.

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.

Hong, Guanglei, & Yu, Bing. (2008). Effects of kindergarten retention on children's social-emotional development: an application of propensity score method to multivariate, multilevel data. *Developmental Psychology, 44*(2), 407.

Hoyle, Rick H. (1995). *Structural equation modeling: Concepts, issues, and applications*: Sage Publications.

Klein, Stephen, Liu, Ou Lydia, Sconing, James, Bolus, Roger, Bridgeman, Brent, Kugelmass, Heather, . . . Steedle, Jeffrey. (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

Klein, Stephen, Zahner, Doris, Benjamin, Roger, Bolus, Roger, & Steedle, Jeffrey. (2013). Observations on AHELO's Generic Skills Strand Methodology and Findings: Council for Aid to Education.

OECD. (2012). Assessment of Higher Education Learning Outcomes. Feasibility Study Report. Volume1 - Design and Implementation.

OECD. (2013). Assessment of Higher Education Learning Outcomes. Feasibility Study Report. Volume3 - Further Insights.

Ou, Suh-Ruu, & Reynolds, Arthur J. (2010). Grade retention, postsecondary education, and public aid receipt. *Educational Evaluation and Policy Analysis, 32*(1), 118-139.

Puhan, Gautam, Sinharay, Sandip, Haberman, Shelby, & Larkin, Kevin. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education, 23*(3), 266-285.

Raudenbush, Stephen W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34*(5), 25-31.

Rosenbaum, Paul R, & Rubin, Donald B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, Paul R, & Rubin, Donald B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

Wise, Steven L., & DeMars, Christine E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment, 10*(1), 1-17.

Wolf, Raffaela, Zahner, Doris, Kostoris, Fiorella, & Benjamin, Roger. (2014). *A Case Study of an International Performance-Based Assessment of Critical Thinking Skills*. Paper presented at the American Educational Research Association, Philadelphia, PA.

Zahner, Doris, & Steedle, Jeffrey. (2014). *Evaluating Performance Task Scoring Comparability in an International Testing Program.* Paper presented at the American Educational Research Association, Philadelphia, PA.