The AHELO Feasibility Study:

Study Results and the Conclusions of the Technical Advisory Group (TAG)

Peter T. Ewell

October 11, 2013

The purpose of this brief paper is to describe the design and results of a Feasibility Study of the Assessing Higher Education Learning Outcomes (AHELO) initiative conducted by the OECD in 2009-2012, and to draw some conclusions about what was learned through the Feasibility Study from the perspective of its Technical Advisory Group (TAG). The TAG was established early in the Study to provide technical guidance to the AHELO Feasibility Study and, based on this experience, to draw conclusions on the conduct of a possible Main Study.[1]

Background and Early Development of AHELO. OECD's Directorate for Education began planning for an international assessment of learning in higher education in the spring of 2007. An initial "experts meeting" was held in Washington, DC in April of that year and involved a group of invited assessment and policy representatives from a half dozen countries. While discussions at that meeting remained general, it was agreed that such an assessment would have value for participating countries and that OECD should design and implement an exploratory study in a few countries in order to identify the challenges involved. This was followed by larger experts meetings in Paris in early July and in Korea in late October.

By the end of the third meeting, the basic shape and timetable of the AHELO initiative had emerged. The assessment would involve several subject matter domains producing results at the individual institutional level for benchmarking purposes. Like the PISA assessments already offered by OECD, they would prominently feature task-like production measures. Two components of the assessment would examine outcomes in selected fields of study with international relevance. After some discussion, Economics and Engineering were chosen as the primary subjects of interest. Another component of the proposed study would be a direct examination of "generic skills" like critical thinking and problem solving that might apply to all fields. One reason advanced for the latter was that "valid and reliable" instruments were already available in this area such as the U.S. Collegiate Learning Assessment (CLA). The experts recognized that such generic skills are widely viewed as critical for individual success, but that any assessment of them should properly be set in the context of particular fields of study. It was also recognized that many fields like history, literature, and law, are so culturally conditioned that international benchmarking would make little sense. Finally, recognizing the vast differences in the contexts for higher education across countries including curricular structure, educational values, and predominant pedagogies, a final important component would be a "Contextual Dimension." This would consist of background surveys completed by faculty and institutional administrators, as well as a student survey to be administered at the same time as each cognitive assessment.

[1] This paper's author served as the Chair of the TAG and its contents draw heavily on reports of the TAG issued throughout the Feasibility Study's history.

Before establishing the detailed design of the Feasibility Study, additional decisions needed to be made about AHELO's purpose and how it should be governed. With respect to purpose, OECD argued that the need to align higher education outcomes in key areas across boundaries in a time of growing graduate mobility was becoming imperative. This is a major objective of the Bologna Process in Europe, and is also reflected in initiatives like the "Tuning Project," which is trying to coordinate academic standards across institutions in different countries in multiple subjects. Reasons for individual institutions to participate reflect the same motivations as institutions administering national standardized assessments like the CLA: such information can be useful for strategic planning and supporting external quality assurance reviews, as well as for benchmarking performance on locally developed measures of student achievement.

With respect to governance, AHELO is located within OECD in the Institute for Management in Higher Education (IMHE), a unit primarily concerned with providing services to institutions in common areas like strategic planning and assessment throughout the world. Policy within this unit is set by the Education Policy Committee (EdPC), which with the OCED Secretariat, assumed broad responsibility for the AHELO initiative. To provide more detailed guidance to the initiative, a Group of National Experts (GNE) was created consisting of representatives of participating countries and jurisdictions, as well as "observer" countries that expressed interest in the initiative but did not participate in the Feasibility Study. Members of the GNE are a mix of policy and ministry representatives and individuals with expertise in international testing.

The AHELO Feasibility Study was originally envisioned to be a two to three year effort involving a small group of countries. The primary questions that this effort was designed to answer were: a) could valid and reliable assessment instruments be developed in multiple languages across quite different country contexts; b) could such an effort be managed politically through OECD's complex multinational governance structure; c) could the testing process be effectively implemented through the use of country and institutional coordinators, assessment administration manuals, and local training programs and; d) would the resulting data be able to be analyzed and prove useful to participating countries and institutions?

The Design of the AHELO Feasibility Study. Each country participating in the Feasibility Study was asked to select a set of six to ten institutions chosen to reflect the diversity of that country's higher education system. The sampling approach for each institution then involved choosing 200 students who were nearing the end of their three or four year period of tertiary study. Because the principal purpose of the Feasibility Study was to try out an assessment approach and draw conclusions about implementation, the resulting data were never intended to support comparisons across institutions or countries. To achieve maximum efficiency, all assessments were "spiraled" so that different students completed different parts of the whole. All students participating in AHELO completed a ninety minute to two hour assessment instrument in one of the domains of interest, plus the short survey designed to gather information about student backgrounds and educational experiences. All of the assessments were administered electronically through secure computer links. Constructed response task (CRT) items were scored by trained reviewers in each country using specially-designed rubrics.

The core of the Generic Skills component of AHELO was a modified version of the CLA, the task-based assessment administered by the New York based Council on Aid to Education (CAE). Representatives of CAE attended each of the initial planning meetings for AHELO and CAE was chosen as one of the prime contractors because its CLA assessment was at that time the only generic skills assessment that used a constructed response format. Two CLA task prompts were selected for further development after an invitational meeting in New York attended by prospective countries—"Lake to River," which deals with the pros and cons of proceeding with a dam project with uncertain safety consequences and "Catfish," which deals with ascribing the likelihood that an observed biological anomaly might be due to a commercial polluter. In addition to completing one of these tasks, each student also answered a short battery of multiple-choice questions (MCQs) on generic competencies drawn from the Graduate Skills Assessment designed by the Australian Center for Educational Research (ACER).

The AHELO Economics assessment was developed by the Educational Testing Service (ETS) using a combination of existing multiple-choice test items drawn in Economics and various international inventories of economic content knowledge including results of the Tuning project and Subject Benchmarks created for the discipline by the Quality Assurance Agency of the United Kingdom. Construction of the assessment was guided by an assessment framework created by a group of subject matter experts. This group also reviewed the assessment itself after draft items were developed. Rather than concentrating on strict content knowledge, the focus of the assessment was on students' ability to "think above content" and use the concepts and language of the discipline effectively. The assessment itself consisted of a set of constructed-response tasks plus a multiple-choice battery drawn from re-worked items on the Economics GRE.

The Engineering assessment was developed by an international partnership of assessment development organizations led by ACER, and was guided by existing national competency frameworks, a Japanese engineering licensure examination, and the results of the Tuning project in Europe. Because Engineering consists of various sub-fields, a decision had first to be made about which should be examined. The panel of subject-matter experts chose Civil Engineering, largely because it is the most conceptually straightforward. In addition, a number of more general skills in Engineering shared by all sub-fields such as analysis and design, as well as basic scientific concepts underlying all of them, were assessed. Construction of the assessment was also guided by an assessment framework, constructed by the subject matter experts. Like the Economics assessment, a focus of the Engineering assessment goes beyond mastery of content to examine students' ability to "think like an engineer." The design of the assessment itself was influenced heavily by the Engineering licensing examination in Japan. Like the assessment in Economics, the Engineering assessment consisted of a mix of multiple-choice items and constructed response tasks.

Although the original plan for the Feasibility Study called for only five or six countries, a total of seventeen countries eventually participated in twenty-five fields—nine in Generic Skills, seven in Economics, and nine in Engineering. All countries participated in contextual data collection, which involved the ten minute student survey administered in conjunction with the assessment, a survey of faculty, a survey of department chairs (in Economics and Engineering) and a survey of institutional administrators. Additional contextual data included a range of descriptive materials about curriculum

and each country's higher education system assembled by a study coordinator in each country. Testing began in the spring of 2012 and was concluded at the end of May, with scoring and analysis completed by the end of the summer of 2012.

Creation and Role of the TAG. The need for an advisory body responsible for reviewing and upholding technical standards for the AHELO Feasibility Study was recognized by the OECD Secretariat and interested countries from the outset of the AHELO initiative. This need was affirmed by the three expert group meetings held in 2007 in Washington, Paris, and Seoul.

The TAG was formally established in 2010 with eight members drawn from assessment and policy experts from throughout the world. The TAG reported to both the OECD Secretariat and the AHELO GNE. The Terms of Reference of the TAG were specified broadly, but essentially established a role that consisted of a) reviewing draft materials on all aspects of the Feasibility Study and suggesting mid-course corrections, b) providing recommendations on the eventual conduct of an AHELO Main Study and, c) providing a definitive recommendation on the feasibility of AHELO at the conclusion of the study. A fourth responsibility was added in Phase II of the Feasibility Study when the TAG was charged with serving as the expert group for the Generic Skills strand and the Context Dimension. Finally, the Terms of Reference established that the GNE could call on the TAG for advice on technical "or other matters"—a charge that allowed the TAG to consider policy and implementation questions with increasing frequency as the Feasibility Study progressed.

The TAG met eight times in the course of the Feasibility Study, three of which were face-to-face meetings and the balance conducted via teleconference. Most meetings consisted of updates on progress guided by a review of documents and covered all facets of the study including the development of assessment frameworks, instrument development, sampling approaches, country coordination, assessment administration procedures, scoring arrangements for CRTs, analysis plans, and reporting arrangements. Recommendations for mid-course guidance of the Feasibility Study were developed by the TAG in the course of these reviews. After each meeting, the Chair of the TAG drafted a report, which was then forwarded to the GNE and the Secretariat. The Chair of the TAG also met with the GNE after each face-to-face meeting to report on issues and recommendations, and the Chair of the GNE also observed the final face-to-face meeting of the TAG in October of 2012.

The TAG's Overall Assessment of the Feasibility Study. The AHELO Feasibility Study constituted an unprecedented multi-national data collection effort at the higher education level. Data on student learning outcomes were collected in three domain strands in seventeen different countries or systems, using assessment instruments comprising both production-focused CRTs and forced-choice MCQs. Data were collected on a wide range of contextual factors by means of surveys completed by students, faculty members, institutional coordinators and national project managers. Numerous implementation challenges including translation, contextualization, sampling, electronic test administration, CRT response scoring, data cleaning, statistical analysis, and reporting were, for the most part, met and successfully overcome. To be sure, some countries/systems experienced more difficulty than others and, because of this, levels of success varied. Nevertheless, all participating countries reported they learned something from the experience and would do it again. Just as important, the Feasibility Study generated

a range of important findings about student learning at the higher education level, as well as dozens of lessons about how a Main Study should be implemented.

That said, some things went particularly well in the AHELO Feasibility Study and a few did not go so well. Most have implied lessons for any AHELO Main Study.

What Went Well. The following were particular strengths of the Feasibility Study:

- Assessment Administration. Electronic administration of assessment on a global scale, and in multiple languages and jurisdictions, presented the Feasibility Study with an enormous challenge. This challenge was met admirably. Only one significant failure in administration occurred over scores of testing sessions at hundreds of institutions. The technical infrastructure underlying this achievement, the thorough training regimens put in place for institutional coordinators, and the robust administration procedures established all contributed to success here.

- Technical Aspects of the Data Analysis. The data yield of the Feasibility Study was large and complex, resulting from the administration of six different instruments to many different kinds of respondents. In the face of this, efforts to provide sound analyses were exemplary from a technical standpoint. The analysis plans were sound, the statistical techniques employed were proper and well executed, and appropriate and effective "work-arounds" were put into place when analytical problems (such as missing data or malfunctioning items) were encountered.

- Instrument Design for Purpose-Built Instruments. All of the instruments designed especially for the Feasibility Study were of exemplary technical quality including the MCQs and CRTs for Engineering and Economics and the three surveys comprising the Contextual Dimension. All were developed through reference to adequate and helpful Assessment Frameworks and were informed by knowledgeable expert groups (in the cases of Engineering and Economics) or considerable background work (in the case of the Contextual Dimension). Moreover, these instruments were produced quickly with little re-work, were designed to a high technical standard, and were piloted as well as could be expected in the short timelines available.

- Overall Coordination. Management and coordination of an enterprise as complex as the AHELO Feasibility Study involved massive challenges of maintaining consistent procedures across five continents, seventeen unique cultural-political contexts, and numerous time zones. The administrative arrangements that were put in place to run the Study met these challenges with clear direction and minimum confusion. Where the inevitable problems were encountered, they were for the most part resolved quickly and smoothly.

Things that Did Not Go So Well. At the same time, some aspects of the Feasibility Study did not go so well. As a consequence, they constitute areas that must be particularly examined as the initiative moves forward.

- Resources and Time. The AHELO Feasibility Study was seriously under-resourced and was implemented on far too short a timeline. More resources and time could have enabled such important features as more cognitive interviews and pilots of newly-build instruments, full-scale field trials of administration and scoring arrangements, and more time for de-briefing and collective discussion of obtained results.

- Uneven Student Cooperation Rates. The validity of any assessment depends in part on obtaining a sufficient number of students chosen as part of the sample at each institution to participate. Countries and institutions achieved substantially different levels of cooperation in the AHELO Feasibility Study. In some countries (Colombia and Mexico, for example) almost all students completed the assessments. In others (Norway and The Netherlands, for example) so few did so that obtained results were not valid enough to use. This mixed track record means that unusual attention to obtaining needed levels of student cooperation will be needed for a Main Study.

- CRT Difficulty and Contextualization. While the CRTs used by the Engineering and Economics assessments were of high technical quality, they were simply too difficult for many students to effectively engage and perform well. At the same time, the CRTs used in Generic Skills based on the CLA proved excessively "American" in an international context. As above, more time for piloting and field trials might have revealed both of these situations at an earlier stage—in time for them to be rectified.

- Contractual Arrangements. The AHELO Feasibility Study began with separate contracts between the OECD Secretariat and the two principal contractors—ACER and CAE. These independent contractual relationships resulted in poor communication among the contractors and occasional duplication of effort. Furthermore, no tendering process was used to procure or develop instruments for the Generic Skills strand. By the time this situation was addressed by re-structuring contractual arrangements so that CAE was a subcontractor of ACER under the Consortium, past history meant that it was difficult to establish a true culture of partnership.

Some Particular Lessons from the Feasibility Study. Experience with the AHELO Feasibility Study offers additional lessons that should be taken forward for the AHELO Main Study:

- ***There should be more opportunities for stakeholder participation in assessment design and in the analysis of assessment results***. There were many points in the Feasibility Study at which the wisdom of practitioners and the national and institutional levels could have been better collected and used for improvement. While the many efforts to contextualize instruments and administration procedures were admirable and, for the most part, successful, a more collaborative approach might have yielded even greater benefits.

- ***A full-scale try-out of all instruments and administration arrangements could enable stakeholder participation in a "design-build" process that would both pilot these designs and***

enable more stakeholder engagement in making them better. This is especially the case for reporting results and sharing data with countries and institutions. Many institutional participants were somewhat disappointed by the lack of attention to their needs for information resulting from the study. Institution-level reports with more detailed breakdowns across student populations would have been beneficial, as well as more fully documented institutional and country data files. A project-wide Quality Monitor should also be established, as well as a National Quality Monitor for each participating country/system. This is consistent with international standards in conducting such studies.

- ***More information should be made available about the costs and benefits to countries and institutions of participating in AHELO.*** Two primary questions will probably be raised by any country/system considering whether or not to join an AHELO Main Study: "what is it likely to cost us?" and "what are we likely to learn?" Because the AHELO Feasibility Study is only just concluded, little can be said about the second question at this point. But some information about costs is available. The direct monetary costs of developing, adapting, and administering the various instruments are known through OECD contracting records. Many costs incurred by institutions and systems for such activities as sampling, student recruitment, test administration, scoring and data reporting, and coordination/oversight are similarly known. But many are not documented because they constitute less tangible costs, for example the time devoted to AHELO by institutional and system personnel. As a consequence, a systematic effort to collect data about both direct and indirect costs should be included in any future Main Study.

- ***Tools such as "Readiness Criteria" should be developed and put in place to allow potential participants and the OECD to determine whether institutions and jurisdictions can actually undertake and benefit from AHELO.*** An explicit set of country and institutional readiness criteria should be established to govern institutional participation in any AHELO Main Study. These criteria should include the provision of a student population sampling frame, sufficient computing infrastructure and IT personnel to support computer-based testing, commitment to participation in training, and effective internal management. It should also include a formal commitment to carry out Study protocols and to abide by the *AHELO Technical Standards*.

- ***Further work is needed about how to effectively assess Generic Skills in multiple disciplinary and national/cultural contexts.*** A major design choice for the AHELO Main Study is whether or not to include a dedicated Generic Skills strand. The existence of these competencies independent of discipline or field of study is a contested issue in the field of higher education assessment. Some generic competencies transfer relatively well across domains, other generic competencies are developed, applied, and assessed much more appropriately within the contexts of particular domains. Results of the Feasibility Study on Generic Skills CRTs suggest that these tasks might perform better if they were better contextualized. The two Generic Skills CRTs used in the Feasibility Study were contextualized to a "real world" problem-solving situation. However obtained results suggest that the manner in which these tasks were

culturally situated and perceived varied substantially across countries and systems. How appropriate contextualization of Generic Skills should be accomplished in any future AHELO Main Study is still a matter for consideration. One option is to continue down the path of including "discipline-specific generic" components in each disciplinary assessment. This was done in Engineering in the Feasibility Study and, to some extent in Economics. If further development along these lines is pursued, these "discipline-specific generic" competencies should be more appropriately aligned with one another to ensure that they address some parallel content. If a decision is made to continue with a separate Generic Skills strand, the performance tasks might be situated in the context of broad disciplinary groupings like the sciences, social sciences, humanities and fine arts.

- ***To provide meaningful information for improving teaching and learning, a mix of item types is required in international assessments at the higher education level***. Another design choice about instrumentation is whether production based CRTs should be included in an AHELO Main Study at all. Decades of research have shown that CRTs will never perform as well in terms of reliability as a battery based solely on MCQs. Results of the Feasibility Study confirm this conclusion for all three domains. The question for an AHELO Main Study is whether the use of CRTs adds enough validity to be worth this inevitable price in lost reliability. On this question, results of the Feasibility Study in Engineering suggest that some of the most important information that could drive improvements in teaching and learning was obtained through the CRTs. The major drawback of including CRTs is substantially increased costs. If the main purpose of AHELO is held to be instructional improvement, the inclusion of CRTs will undoubtedly increase the usefulness of results. On the other hand, if the main purpose is to provide the most reliable international benchmarks of institutional performance with respect to student learning outcomes, the greater reliability and lower cost of adopting an approach based solely on MCQs may be preferred.

- ***An acceptable response rate based on an accurate probability sample is required to assure comparability of results across institutions.*** In the Feasibility Study, probability sampling or a census of students was used by almost three-quarters of participating institutions. For the remaining institutions, it is not apparent why such a sample was not used. For the Main Study, participating institutions should be required to compile a list (or lists) of eligible students (or groups of students) and to use either probability sampling or a census. That said, there should be some flexibility regarding the choice of probability sampling method. For example, cluster sampling of class groups may be reasonable when the number of eligible students is large. It may also be reasonable for AHELO to impose a fixed minimum response rate threshold, at the level of the country or the institution, below which data will be excluded from the data analysis. Finally, measures to increase response rates should be actively researched before any new AHELO data collection.

- ***AHELO should be better located and integrated with the international scholarly community examining student learning outcomes and the policies and practices that support better***

learning. The past decade has seen a sharp increase in policy and scholarly interest in improved academic performance in higher education. Evidence of this can be seen in the Bologna Process and Tuning in Europe, the Spellings Commission and interest in accreditation in the U.S., the rise of qualifications frameworks in many nations, and the emergence of multinational ranking initiatives like U-Map and U-MultiRank. AHELO represents an opportunity to better align the emerging scholarly and policy dialogue about quality.

- ***All of this will require more time and adequate resources***. The AHELO Feasibility Study experienced serious resource shortfalls which, in the course of implementation, negatively affected many of its components. This occurred incrementally and its effects were complicated by the fact that the project included more countries than a "feasibility study" should probably have included. A similar under-resourced condition cannot be allowed for a Main Study. The OECD and participating countries will need to ensure adequate resources in moving forward. If this cannot be guaranteed, implementation will have to wait until it can.

Moving Forward? The OECD has concluded that the results of the Feasibility Study were sufficiently positive that a Main Study should be conducted and has already distributed a paper describing its main features and inviting country participation. This verdict by no means assures that AHELO will become more broadly operational, however. The real question that will govern moving forward is not whether or not it is possible to conduct these assessments. It is instead whether or not the effort is cost effective. The cost side is by this point readily apparent. The AHELO Feasibility Study cost more than nine million Euros to implement, most of which was borne by participating jurisdictions and institutions. Benefits to participants, on the other hand, have been mixed—largely depending upon the amount of effort jurisdictions and institutions invested in undertaking local data analyses and disseminating the results. Consistent with OECD's hopes for a Main Study, participating institutions may find internationally benchmarked assessment results helpful in their strategic planning efforts. Participating countries, meanwhile, will be able to rehearse how various kinds of results can be used to evaluate higher education performance. Whether either of these actors, as well as a host of non-participating institutions and nations, will come to believe that full implementation is worth the investment remains to be seen.