Assessment of Higher Education Learning Outcomes Feasibility Study Report

Volume 2 – Data Analysis and National Experiences

Executive Summary

*This **second volume** will be followed by a third and final volume on Further Insights (including the March Conference Proceedings) at the end of April.*
The first volume on Design and Implementation was published in December 2012.

The complete Volume 2 can be found on the AHELO website (www.oecd.org/edu/ahelo)
http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume2.pdf

Chapter 7 - Validity and reliability insights on scientific feasibility from the AHELO feasibility study data

This chapter was prepared on the basis of the information available at the time of publication. However the unavailability of certain information did not allow OECD analysts or external experts to replicate or complement the information and analyses the OECD has received. Also, because of the unavailability of some of the psychometric results, the inclusion of the associated conclusions in this report does not imply the OECD's endorsement of the conclusions.

The scientific feasibility of AHELO rests on its capacity to produce valid and reliable results across different countries, languages, cultures and institutional settings. This chapter presents an overview of the data collected and analyses conducted in order to assess the scientific feasibility of the instruments that were used. These analyses and results presented should be interpreted in the "proof of concept" spirit of the feasibility study.

Validity and reliability concepts for assessing scientific feasibility

Validity is a broad concept that involves making appropriate interpretation and uses of test scores. It requires that the purpose and inferences to be drawn from test scores be stated from the outset. The evaluation of instrument validity requires the collection of a variety of evidence to support different types of validity.

Reliability means that test results are consistent and stable across different testing situations. An instrument's degree of reliability can be affected by a number of different factors. Stable results suggest that the observed student scores are more likely to reflect true scores. Reliability of an instrument is classically expressed as the ratio between the true variance, i.e. the true ability, and the observed variance, i.e. the observed test scores that include random factors.

Evidence on scientific feasibility collected during the AHELO feasibility study

Some individual test items may turn out to perform poorly and must be removed before validity and reliability can be assessed. Items not meeting psychometric standards are deleted from final analyses. Non-functioning items can be removed on a country basis.

The small number of items removed from the generic skills instrument indicates that it has good overall item quality while the relatively small number of items deleted from the economics and engineering instruments indicates they have sufficient overall item quality.

In an international study, it is critical to ensure that items have a similar level of difficulty across the different countries. Differential item functioning (DIF) analyses are conducted to further understand differences in performance of different student sub-populations. Most items showed no significant differences in performance between genders. For those who did, further analyses would be needed to identify the underlying reasons for these differences by gender.

Three institutional characteristics are used as a basis for comparison across the different types of higher education institutions. Results indicate difference in student performance across the different institution types, depending on the basis for comparison.

- Little difference in student performance is observed when using the institution size (small/medium/large).
- Differences in student performance are observed when compared on the basis of highest degree the institution offers (baccalaureate, master and doctorate).
- Differences in student performance are observed when using the institution emphasis on research and teaching (research, teaching, and research/teaching balance).

Many items do not function as expected for some countries. Further analyses are needed to identify the underlying reasons for these country differences. Many items do not function as expected for some languages. Constructed-response tasks in the generic skills strand show significant differences in student performance across languages. Further analyses are needed to identify the underlying reasons for **student** performance differences in the different languages.

Validity evidence

Different types of evidence are collected throughout the feasibility study to determine the validity of instruments used.

The three assessment instruments display reasonable levels of construct validity evidence. Results indicate that the overall scale could be divided into complementary sub-scales.

Expert consensus provided evidence of content validity of the economics and engineering instruments but was not fully demonstrated for the generic skills instrument. Feedback from the generic skills cognitive labs showed that the constructed-response tasks were attractive to students. Students also reacted positively to the draft economics and engineering constructed-response tasks. Further content validity evidence for the two discipline instruments is still required to fully confirm content validity.

Face validity is assessed through several indicators.

- Students spent a good deal of time responding to the AHELO assessments. The low levels of non-response indicate good levels of student engagement with the instruments.
- Student reported putting a good deal of effort into the AHELO assessments. Self-reported effort by field of education for students participating in the generic skills strand also reveals limited variations across fields.
- Students' perceptions of the educational and professional relevance of the instruments vary across strands and also reveal some differences across fields of education.

Two indicators are used as criteria to provide concurrent validity evidence:

- Results show a correlation between students' AHELO test scores and their self-reported academic performance only for the engineering strand. The strength of the relationship between AHELO scores and self-reported academic performance varies across countries.
- The strength of the relationship between AHELO scores and students' overall education satisfaction varies across countries.

Reliability evidence

The feasibility study produced instruments with "acceptable" to "good" levels of overall reliability. Examination of reliability indices at the country level shows less reliable results for some of them.

Reliability analyses using data aggregated at the institutional level suggest "acceptable" to "good" levels of reliability in all three strands. Examination of reliability indices using data aggregated at the institutional level indicates less reliable results for some countries.

Inter-scorer reliability statistics for constructed-response tasks can be considered "fair" to "good" in all three strands. Scoring of student responses may vary across countries but their rank ordering is very consistent. Scoring of student responses is consistent across countries when considering the tasks total scores.

The correspondence between the item difficulty levels and the students' ability levels for the generic skills strand indicates that the instrument is well targeted to the student population. The distribution of student performance shows that the economics and engineering tests were too difficult. The large proportion of "zero" scores for the economics and engineering constructed-response tasks also indicates that the items were too challenging for students.

Effort seems to have a greater impact on constructed-response tasks than on multiple-choice items.

Conclusions

Overall item quality and functioning

The AHELO feasibility study produced many items that functioned well.

Overall assessment of validity

All three instruments have achieved reasonable levels of construct validity. The evidence collected also suggests that the instruments have achieved reasonable levels of content validity in the disciplinary strands and suggests that the instruments have achieved reasonable levels of face validity in all three strands. Evidence on concurrent validity is less conclusive.

Overall assessment of reliability

The three instruments provided reliable results. Inter-scorer reliability can be considered "fair" to "good" in all three strands.

Overall scientific feasibility

The AHELO feasibility study demonstrated that it is feasible to develop instruments with reliable and valid results across different countries, languages, cultures and institutional settings.

Chapter 8 – National Experiences

Seventeen countries/economies took part in the AHELO Feasibility Study. We have asked them to reflect on the experience. Their feedback is provided in Chapter 8 of the Feasibility Study Report, country by country. The first page of each country's contribution to the Report is the poster which was prepared for the AHELO feasibility study Conference. These posters are reproduced below.

Abu Dhabi

ABU DHABI

- ☐ Economics
- ☑ Engineering
- ☐ Generic Skills

Abu Dhabi's vision to become a knowledge-based economy depends largely on the quality of its graduating higher education students, and the AHELO feasibility study represents a catalyst in being able to assess their knowledge and skills using a reliable, multifaceted, and internationally valid measure.

Main Challenges

- ✓Short timelines: the development of a communication and recruitment strategy had to be implemented in two weeks, the mobilisation of all stakeholders in less than 1 month and the set up of the national infrastructure in less than two weeks.
- ✓Provision of training on the technical standards and operations guidelines, and for documenting processes to be aligned with 2 years of work accomplished by other countries.
- ✓Recruitment of all in-scope students for the engineering strand on a census basis.

Main achievements

- ✓Assessment Culture: Promoted the culture of comparative assessment at the institutional level and added a new dimension to the QA framework in Abu Dhabi.
- ✓Learning Outcomes: measured the learning outcomes of Abu Dhabi students in comparison to their international peers and informed institutional decision makers and leaders of the capacity of their education systems.
- ✓Capacity Building: developed the capacity of a highly skilled team to lead and implement similar national projects in the future and paved the way for adopting similar assessment exercises at the national level.

Main Lessons

- ✓Success in implementing the project activities smoothly due to the rigorous selection criteria for the ICs, Scorers and TAs.
- ✓The high response rate emphasized the importance of having a well designed communication strategy to explain the expected outputs of AHELO.
- ✓More time should be allowed for scorer recruitment, training and activities.
- ✓Exchange of information and collaboration between countries (on scoring or risk management) illustrated that AHELO would benefit from a larger institutionalized process involving all countries.

Australia

AUSTRALIA

☐ Economics
☑ Engineering
☐ Generic Skills

Participation in AHELO has shown that Australia is well equipped to participate in this type of international study, with interest from stakeholders in participating, and well established systems for implementation.

Main Challenges

- ✓ Motivating students to participate.
- ✓ Securing a representative sample.
- ✓ Highlighting that the main outcomes from this study relate to the processes generated and not the data.

Main achievements

- ✓ Beginning a conversation about learning outcomes in higher education and offering a tool for assessing them.
- ✓ Co-ordination and co-operation of institutions in the implementation of a large online assessment.
- ✓ Stimulating students through innovative forms of assessment.

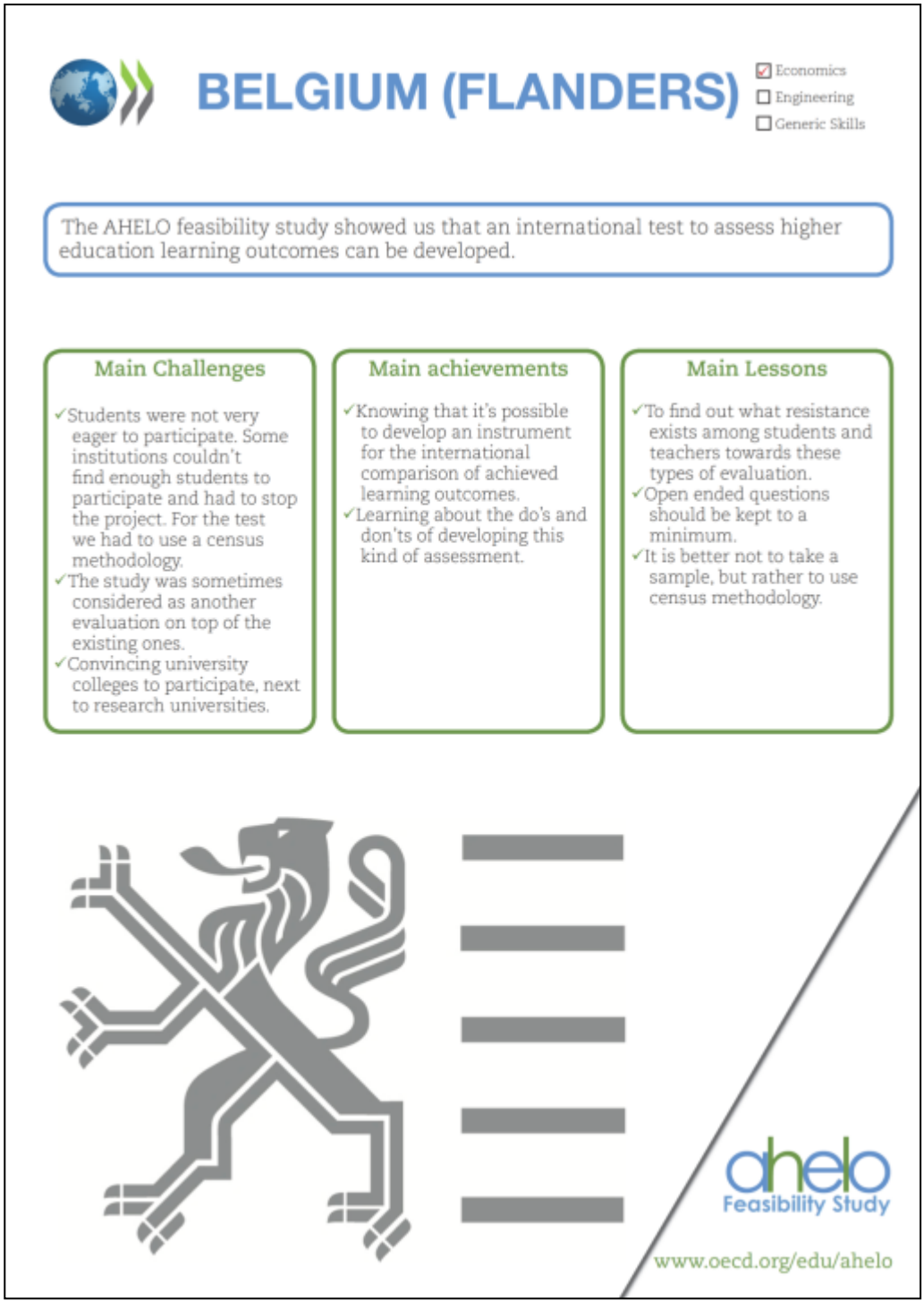
Main Lessons

- ✓ Long term planning is key to successful student engagement in such assessments.
- ✓ There are very motivated and generous people in Australian higher education.
- ✓ Providing ongoing information, data and outcomes summaries is important for maintaining interest and motivation within institutions.

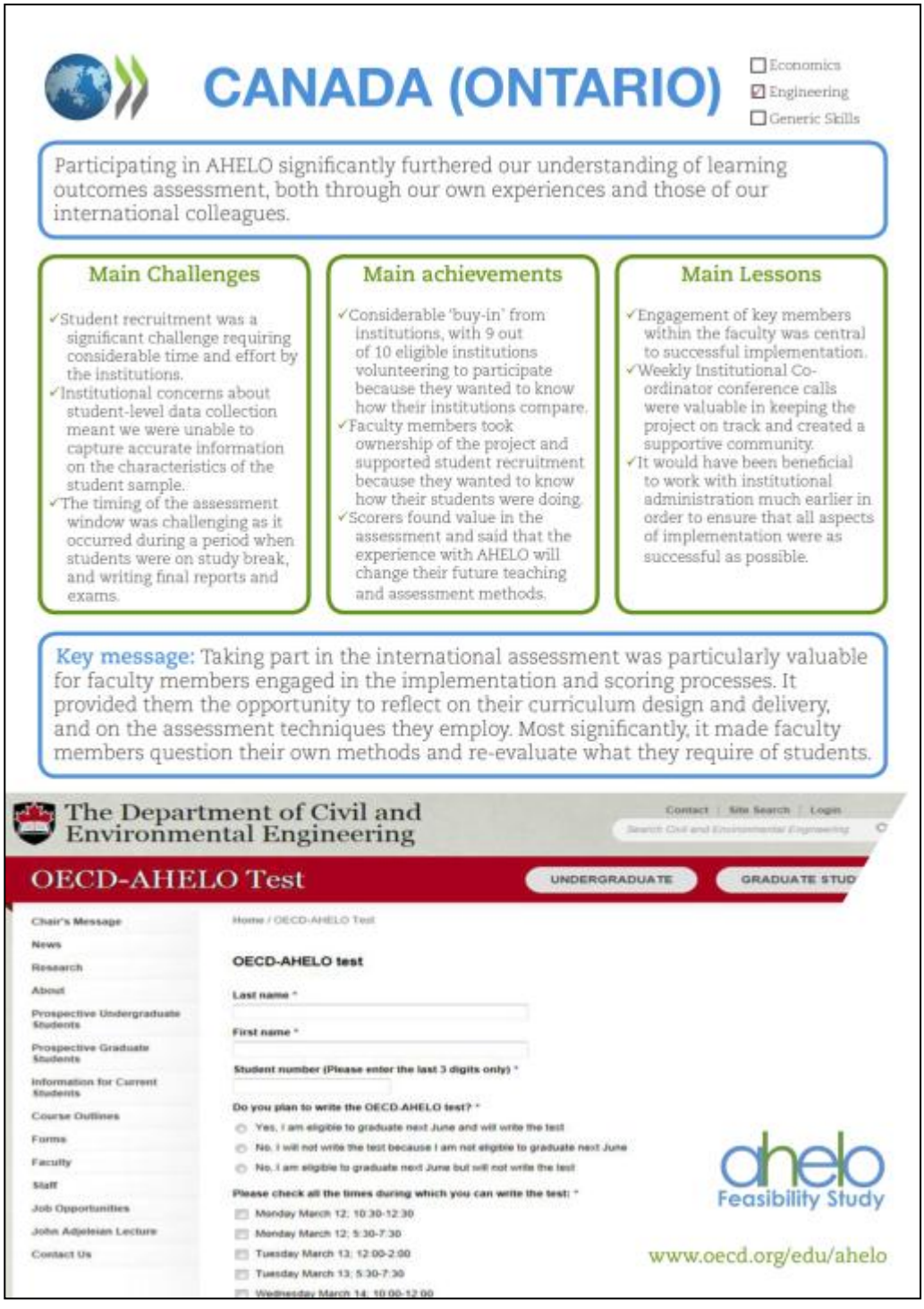
Key message: Australia is proud to be at the forefront of the development and implementation of new ways of evaluating quality and improving learning in higher education.

Australian institution co-ordinators at AHELO Symposium, October 2011

Belgium - Flanders



Canada (Ontario)



Colombia

Egypt

EGYPT

☑ Economics
☑ Engineering
☑ Generic Skills

In light of the inspiring 25th of January revolution, the Egyptian people have expressed their desire for more effective reform, as well as greater expectations for better quality of service in all aspects of life, particularly education. The new era of democracy and transparency is in harmony with concepts such as self-assessment and the developments that a ground breaking reform project like AHELO targets.

Main Challenges

- ✓The rapid and radical changes that involved the whole Egyptian community.
- ✓The repeated changes in the leadership of higher education institutions and management boards that altered the implementation schedule for the project's activities.
- ✓The large Egyptian universities (80 000-250 000 undergraduate students) and their incomplete electronic databases.

Main achievements

- ✓High response rates for students (total number 4 212) and faculty (total number 877), representing 18.3% (students) and 18.2% (faculty) of total AHELO participation.
- ✓Increased awareness of the academic societies regarding the importance of linking the intended outcomes of programmes with the labour market.
- ✓Success of the first concurrent online testing in the participating universities.

Main Lessons

- ✓National governmental commitment and support are cornerstones for assuring success of such large scale research studies.
- ✓Recruitment of students for participation in future studies entails innovative strategies.
- ✓Test simulations using released test instruments should be considered, for training purposes, as well as for exploring pitfalls and how they can possibly be avoided.

Finland

FINLAND

☐ Economics
☐ Engineering
☑ Generic Skills

AHELO gives important information to faculties, institutions and governments on how to develop educational activities to further promote students' learning in the era of globalised higher education.

Main Challenges

- ✓internationally unstable financial situation
- ✓low participation rate of Finnish students, which jeopardises the whole idea of AHELO in helping HEIs to develop their teaching and learning activities
- ✓tight schedule in the implementation phase

Main achievements

- ✓high interest in AHELO among Finnish higher education institutions
- ✓completion of instrument development and implementation in the given timeframe
- ✓firm governmental support and competent national organisation

Main Lessons

- ✓International financing of the project has to be fully secured before it can start.
- ✓The international consortium, including all of the partners, has to have a solid and consistent understanding of what kind of instruments to develop and how to carry out a large-scale international comparative project such as AHELO.
- ✓The implementation phase must be given enough time; more time is needed to motivate students, to train ICs and to organise test sessions.

Key message: In order to avoid problems in the management and steering of AHELO, it is of crucial importance to secure full financing for the project based on a realistic budget. This would also help the participating countries to plan and finance their activities more precisely and ultimately yield a more coherent and manageable project.

Italy

Japan

JAPAN

☐ Economics
☑ Engineering
☐ Generic Skills

For Japan, the AHELO feasibility study represented an exciting engagement in an international conversation on what engineering graduates are expected to know and be able to do in a knowledge based global society.

Main Challenges

- ✓Involving faculties: taking faculty time away from teaching and research requires a very good reason, as well as a clear description of what kind of feedback they will be receiving.
- ✓Sustaining momentum: the prolonged planning period made it difficult to keep the higher education community interested and engaged.
- ✓Translating instruments according to protocols: achieving substantive equivalence requires some flexibility and extensive knowledge of the language and the subject matter.

Main achievements

- ✓A tangible and substantive understanding of a conceptual framework of engineering competencies and learning outcomes that can be shared globally.
- ✓Concrete and innovative ideas for conceptualising and measuring competencies and learning outcomes.
- ✓A delightful experience working on an international team, learning from global partners, and being able to make unique contributions.

Main Lessons

- ✓An international assessment of higher education learning outcomes can become a useful tool for educators to globally benchmark and update their teaching practices.
- ✓Designing constructive response tasks to "measure" how students can "think" like an engineer requires a thoughtful balance between open-endedness and preciseness.
- ✓The exercise of scoring and modifying scoring rubrics by an international team of experts is extremely important to reach consensus on the scope and level of expected learning outcomes.

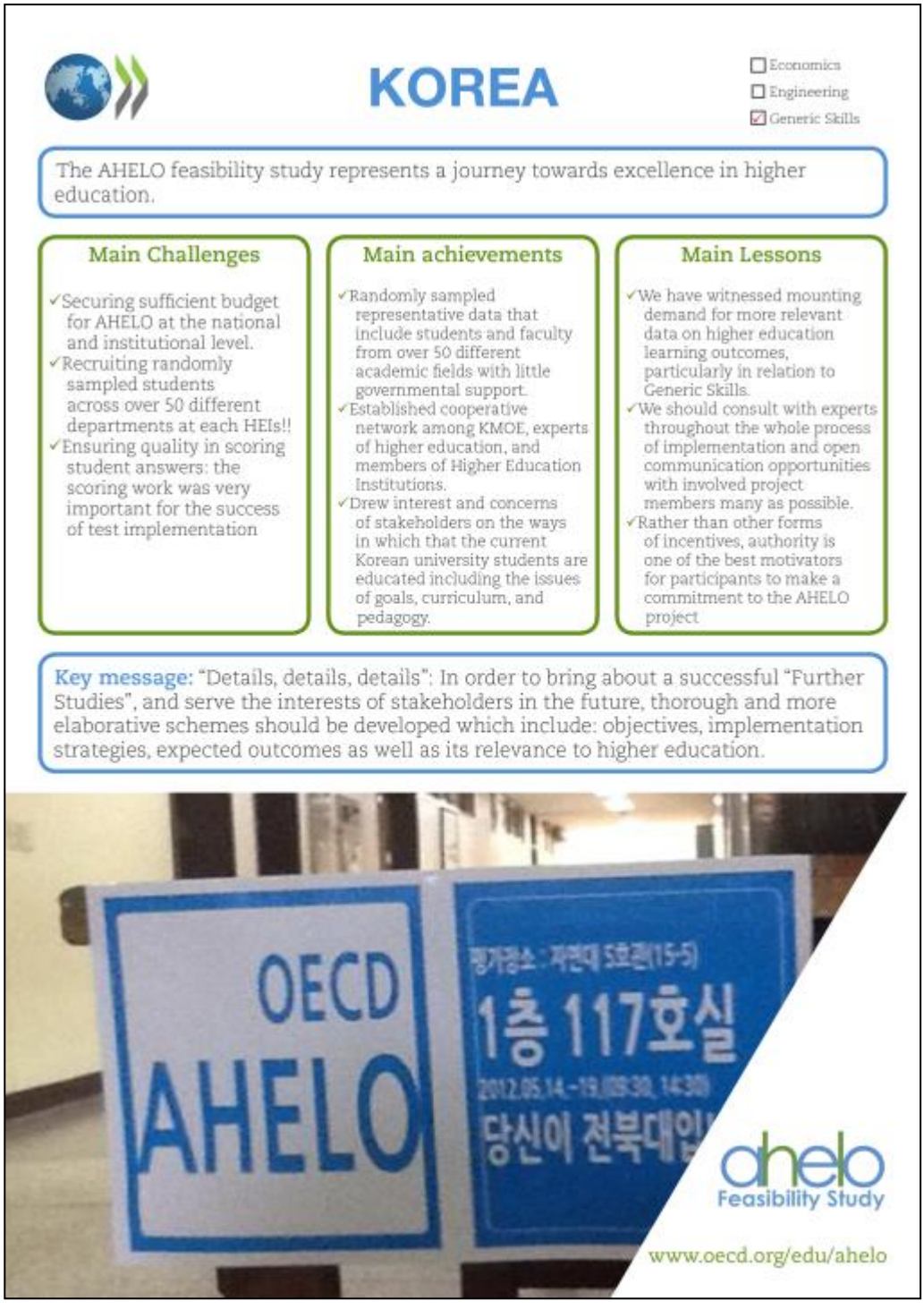
Key message: AHELO can become a powerful tool for educational improvement, when instruments and scoring rubrics are made fully available to participating institutions, and when coupled with workshops that induce discussion about curriculum design and encourage innovation in teaching and learning.

The Japanese Scoring Team (June 2012)

ahelo Feasibility Study

www.oecd.org/edu/ahelo

Korea



Kuwait

Mexico

MEXICO

☑ Economics
☑ Engineering
☑ Generic Skills

Building new institutional knowledge and more assessment capacities for improvement, in a higher education system characterised by wide diversity, constant growth and major challenges.

Main Challenges

- ✓Participation of 2 472 Mexican students in the generic skills strand; 825 in the Engineering strand; and 541 in the Economics strand; with a very good response rate: 75, 82 and 80% respectively.
- ✓High interest, commitment and enthusiasm from authorities, faculty and students of 14 HEIs in Mexico (http://ahelo.uaslp.mx/); participation of nearly a thousand faculty and 150 persons as institutional co-ordinators by strand, test administrators and lead scorers.
- ✓Integration of a Mexican co-ordinating team (with three leading universities: UASLP, UDG and UADY) with excellent levels of collaboration, communication and trust between NPMs, the GNE representative and authorities of Secretaría de Educación. Specifically the Dirección General de Educación Superior Universitaria of the Subsecretaría de Educación Superior: provided advice and substantial financial support.

Main achievements

- ✓Mexican HEIs consider this evaluation approach as an innovative tool for taking decisions and improving the quality of education, since it focuses not just on inputs and processes, but also on outputs that can be referred to contextual variables and international scope.
- ✓We must go deeper into the assessment frameworks and their complex methodological implications, based on a more collegial deliberation among the international, national and institutional levels.
- ✓To achieve the goals and objectives, it is very important to work as a team, with the participation of experts from several disciplines (educators, specific strand experts, psychometric staff, etc.).

Main Lessons

- ✓Maintaining a good rhythm of participation and communication with international working groups and performing all technical tasks with the high quality standards required, in the three strands.
- ✓Motivating the students to participate in the evaluation of their learning competences, particularly for the constructed response sections; although students were not familiar with this type of test, they showed their appreciation for it.
- ✓Addressing the concerns about misusing the results of AHELO feasibility study and taking the proper preventive measures.

Figure 1. Geographical distribution of Mexican universities that participated in the OECD AHELO feasibility study

ahelo Feasibility Study

www.oecd.org/edu/ahelo

Netherlands

NETHERLANDS

- [x] Economics
- [ ] Engineering
- [ ] Generic Skills

The AHELO feasibility study in the Economics Strand proved to be an interesting addition to the ways in which the quality of education is measured in the Netherlands, but challenging to organise on top of all other quality measures already being implemented (none of which allow for an international comparison though).

Main Challenges

- ✓Getting HEIs and students to participate without incentive, despite communication efforts and faculty involvement.
- ✓Realising an international assessment in a very short timeline due to the uncertainty of the project and in a period when Faculty was very busy with other activities.
- ✓Adapting the assessment of the economics strand to the Dutch economic curricula and the Dutch binary system.

Main achievements

- ✓Both types of Higher Education in the Netherlands represented in the field work (Research Universities and Universities of Applied Sciences).
- ✓Despite the very short timeline a successful central management and implementation plan was set up.

Main Lessons

- ✓The items in the economics strand are more knowledge-based than would be expected and they did not cover the whole range of economic studies provided in the initial economics framework.
- ✓In binary systems the development of assessment instruments should cater for the needs of different types of economics studies, in such a way that both research-oriented as well as applied-oriented types of education clearly recognise their own academic content and didactical approach.
- ✓Incorporate the instruments more into the quality assurance measures already being implemented in Higher Education on a national and institutional level.
- ✓Incorporate the main study instruments more into the academic programmes students participate in, or make sure that there is a clear (and equal) incentive offered to all students participating.
- ✓The timing of the assessment should be considered.
- ✓Make sure that in all parts of the academic community involved (i.e. Ministry, national organisations, institutions, faculties, teachers and students) there is enough support for an AHELO and carefully position NC and NFM in this field.

ahelo
Feasibility Study

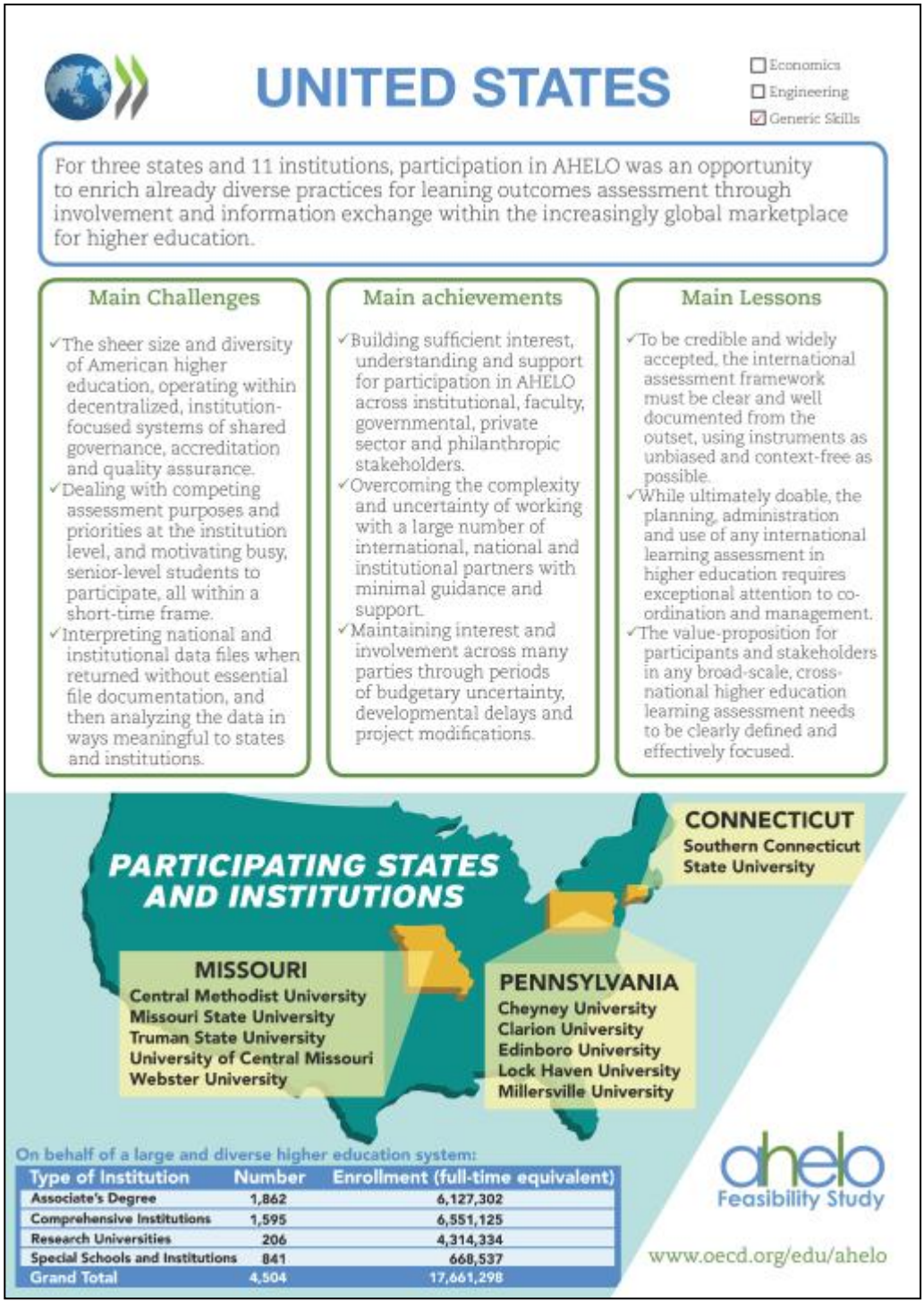
www.oecd.org/edu/ahelo

Norway

Russian Federation

Slovak Republic

United States



Conclusions from Chapter 9 (by Peter T. Ewell, Chair of the TAG)

The TAG's overall assessment of the feasibility study

The Technical Advisory Group (TAG) believes that the AHELO feasibility study constituted an unprecedented multi-national data collection effort at the higher education level. Data on student learning outcomes have been collected in three domain strands in 17 different countries or systems, using assessment instruments comprising both production-focused CRTs and forced-choice MCQs. Data have also been collected on a wide range of contextual factors by means of surveys completed by students, faculty members, ICs and NPMs.

Numerous implementation challenges including translation, contextualisation, sampling, electronic test administration, CRT response scoring, data cleaning, statistical analysis, and reporting have been 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 most 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 such a project should be implemented in the future.

That said, the TAG wishes to briefly point out a few things that went particularly well in the AHELO feasibility study and a few that did not go so well. Several of these have been touched upon in earlier sections of the report and most have implied lessons for any AHELO Main Study.

What went well

The TAG believes that 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, confronted 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 Co-ordinators, and the robust administration procedures established were all praiseworthy.

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, the Consortium's 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 co-ordination

Management and co-ordination of an enterprise as complex as the AHELO feasibility study involved massive challenges of maintaining consistent procedures across five continents, 17 unique cultural-political contexts, and numerous time zones. The administrative arrangements established by the Consortium 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, the TAG believes that some aspects of the feasibility study did not go so well. As a consequence and as reflected in the TAG's recommendations for any AHELO Main Study, they constitute areas that must be particularly examined as the initiative moves forward.

Resources and time

As the TAG pointed out repeatedly in the course of the feasibility study, 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.

CRT difficulty and contextualisation

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 it to be rectified.

Reporting results

While the TAG believes that the Consortium's analyses of the massive amount of data generated by the feasibility study were exemplary from a technical standpoint, the reporting of these results through the Consortium's final report was overly complex, and therefore difficult to understand. Most important, the report lacked clearly stated conclusions on which to make policy decisions for the future. Again, this was probably partly a result of time pressures, and

the reporting process would have benefitted from reflection and feedback from stakeholders after results were made available. Again, the March 2013 conference should prove useful in this respect.

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 — a fact that is highly unusual in international studies of this kind. By the time this situation was addressed by re-structuring contractual arrangements so that CAE was a subcontractor of ACER under the Consortium, a habit of independence — exacerbated by commercial rivalry—made it difficult for both parties to establish a culture of partnership.

Some additional lessons

Finally, the TAG believes that the AHELO feasibility study offers several additional lessons that should be taken forward for any international assessment effort of this size and scale:

- ***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 contextualise instruments and administration procedures were admirable and, for the most part, successful, a more collaborative approach might have yielded 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 NPMs and ICs remain somewhat disappointed by the lack of attention to their needs for information resulting from the study — especially the provision of country-level data files that lacked the documentation needed for analysis.

- **Any such study should be better located and integrated with the international scholarly community examining student learning outcomes and the policies and practices that support better learning.** As pointed out in the rationale for AHELO, the past decade has seen a sharp increase in policy and scholarly interest in improve 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 United States, the rise of qualifications frameworks in many nations, and the emergence of multinational mapping and 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 TAG's conclusion in this regard remains unaltered: if the required resources and timelines needed are not forthcoming, a future study of this kind should not be undertaken.

On balance, the TAG believes firmly that the AHELO feasibility study was soundly executed and provided many lessons that will continue to inform international assessment efforts for many years to come. Among its most important contributions to the study were recommendations to ensure consistency of administration and scoring across contexts, steady reinforcement of the need for contextual data — especially at the beginning of the study, recommendations to reinstate an MCQ component in Generic Skills, and recommendations to the OECD Secretariat about how to prepare its final report. Members of the TAG all learned something important through their engagement in the study and congratulate the Consortium and the OECD Secretariat for a job well done.