

ASSESSMENT OF HIGHER EDUCATION LEARNING OUTCOMES

AHELO

FEASIBILITY STUDY REPORT

VOLUME 2

DATA ANALYSIS AND NATIONAL EXPERIENCES



Assessment of Higher Education Learning Outcomes

Feasibility Study Report

Volume 2 – Data Analysis and National Experiences



This work is published on the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Organisation or of the governments of its member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgement of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to rights@oecd.org. Requests for permission to photocopy portions of this material for public or commercial use shall be addressed directly to the Copyright Clearance Center (CCC) at info@copyright.com or the Centre français d'exploitation du droit de copie (CFC) at contact@cfcopies.com.

ACKNOWLEDGEMENTS AND CREDITS

Chapter 7 of this report has been written by Karine Tremblay, Diane Lalancette and Deborah Roseveare of the OECD Directorate for Education.

Chapter 8 was prepared by the National teams of the countries/economies who participated in the AHELO feasibility study.

Chapter 9 was prepared by Peter Ewell, the Chair of the AHELO Technical Advisory Group.

As with the first Volume of this Feasibility Study Report the Secretariat also wishes to acknowledge the extensive contributions from the AHELO Consortium report authors and contributors: Hamish Coates, Sarah Richardson, Yan Bibby, Stephen Birchall, Eva van der Brugge, Rajat Chadha, Steve Dept, Jean Dumais, Daniel Edwards, Thomas van Essen, Peter Ewell, Andrea Ferrari, Jacob Pearce, Claire Melican, Xiaoxun Sun, Ling Tan, Rebecca Taylor and Don Westerheijden.

Special thanks are due to Emily Groves who edited the first drafts of the country contributions, as well as to Cécile Bily who edited the final version, drafted the introduction, and readers' guide and prepared this report for publication.

Thanks are also due to the many other OECD colleagues who contributed to this project at different stages of its development including Barbara Ischinger, Andreas Schleicher, Richard Yelland, Fabrice Hénard, Valérie Lafon and Sabrina Leonarduzzi. The AHELO feasibility study also benefited from the contributions of the following consultants, seconded staff and interns: Rodrigo Castañeda Valle, HoonHo Kim, Claire Leavitt, Eleonore Perez Duarte, Alenoush Saroyan, Tupac Soulas, Takashi Sukegawa and Mary Wieder.

The Secretariat would also like to express its gratitude to the sponsors who, along with the participating countries, generously contributed to this project and without whom the AHELO feasibility study would not have been possible: Lumina Foundation for Education (USA), Compagnia di San Paolo (Italy), Calouste Gulbenkian Foundation (Portugal), Riksbankens Jubileumsfund (Sweden), the Spencer and Teagle Foundations (USA) as well as the higher Education Founding Council – HEFCE (England) and the Higher Education Authority – HEA (Ireland). The William and Flora Hewlett Foundation also provided support for U.S. participation in the study.

And finally a special word of thanks to Jan Levy, the Chair of the AHELO GNE, who provided invaluable guidance and support to the Secretariat throughout the feasibility study.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS AND CREDITS	3
INTRODUCTION	7
READERS' GUIDE	8
Note on terminology	8
Abbreviations used in the feasibility study report	8
CHAPTER 7 VALIDITY AND RELIABILITY	
INSIGHTS ON SCIENTIFIC FEASIBILITY FROM THE AHELO FEASIBILITY STUDY DATA	11
Validity and reliability concepts for assessing scientific feasibility	12
Evidence on scientific feasibility collected during the AHELO feasibility study	14
Conclusions	30
REFERENCES	33
CHAPTER 8 NATIONAL EXPERIENCES	39
Abu Dhabi	40
Australia	48
Belgium - Flanders	54
Canada (Ontario)	60
Colombia	66
Egypt	72
Finland	82
Italy	86
Japan	90
Korea	96
Kuwait	102
Mexico	112
Netherlands	120
Norway	126
Russian Federation	132
Slovak Republic	140
United States	146
CHAPTER 9 ROLE OF THE AHELO FEASIBILITY STUDY TAG	153
Creation and Role of the TAG	154
Major TAG recommendations made during the Study	155
Recommendations on the conduct of an AHELO Main Study	163
The TAG's overall assessment of the feasibility study	168

REFERENCES	172
ANNEX C – ADDITIONAL TABLES AND FIGURES	175
ANNEX D– TAG TERMS OF REFERENCES	193

Tables

Table 7.1 - Number of items removed by country - Generic Skills	15
Table 7.2 - Number of items removed by country - economics	16
Table 7.3 - Number of items removed by country - engineering	17
Table 7.4 - Item functioning differences across countries	22
Table 7.5 - Item functioning differences across languages	23
Table C1 - Generic skills assessment multiple-choice items item-level non-response by item rotation (n=10657)	175
Table C2 - Economics assessment item-level non-response by item group (n=6242)	175
Table C3 - Engineering assessment item-level non-response by item group (n=6078)	175
Table C4 - Generic skills cognitive labs follow-up questions	176
Table C5 - Student feedback collected during focus groups (economics and engineering strands)	176
Table C6 - Overall instrument reliability estimates, by strand by countries	177
Table C7 - Generic skills inter-scorer reliability statistics (n=10657)	177
Table C8 - Economics scoring inter-scorer reliability statistics (n=8325)	178
Table C9 - Engineering inter-scorer reliability statistics (n=8084)	178

Figures

Figure 7.1 - Gender DIF analyses	18
Figure 7.2 - Institution DIF analyses: Size (small versus medium)	19
Figure 7.3 - Institution DIF analyses: Size (medium versus large)	19
Figure 7.4 - Institution DIF analyses: Highest degree offered (doctorate versus masters)	20
Figure 7.5 - Institution DIF analyses: Highest degree offered (doctorate versus baccalaureat)	20
Figure 7.6 - Institution DIF analyses: Emphasis (research versus teaching)	21
Figure 7.7 - Institution DIF analyses: Emphasis (research versus research/teaching balance)	21
Figure C1 - Self-reported effort put into the generic skills assessment, by country (n=10657)	179
Figure C2 - Self-reported effort put into the economics assessment, by country (n=6242)	179
Figure C3 - Self-reported effort put into the engineering assessment,	

by country (n=6078)	180
Figure C4 - Self-reported effort put into the generic skills assessment, by field of education (n=10657)	180
Figure C5 - Student perceptions of relevance of the assessment instrument, by strand	181
Figure C6 - Student perceptions of relevance of the generic skills assessment, by field of education (n=10657)	181
Figure C7 - Generic skills score and self-reported academic performance, by countries (n=10657)	182
Figure C8 - Economics score and self-reported academic performance, by country (n=6242)	182
Figure C9 - Engineering Score and self-reported academic performance, by country (n=6078)	183
Figure C10 - Generic skills scores and overall education satisfaction, by countries (n=10657)	183
Figure C11 - Economics score and overall education satisfaction, by country (n=6242)	184
Figure C12 - Engineering score and overall education satisfaction, by country (n=6078)	184
Figure C13 - Generic skills assessment variable map (n=10657)	185
Figure C14 - Economics assessment variable map (n=6242)	186
Figure C15 - Engineering assessment variable map (n=6078)	187
Figure C16 - Economics assessment zero scores, by CRT and country (n=6242)	188
Figure C17 - Engineering assessment zero scores, by CRT and country (n=6078)	188
Figure C18 - Generic skills score variance explained by effort, by country and task type (n=10657)	189
Figure C19 - Economics score variance explained by effort, by country and task type (n=6242)	190
Figure C20 - Engineering score variance explained by effort, by country and task type (n=6078)	191

INTRODUCTION

In 2008, the OECD launched the AHELO feasibility study, an initiative with the objective to assess whether it is possible to develop international measures of learning outcomes in higher education.

Learning outcomes are indeed key to a meaningful education, and focusing on learning outcomes is essential to inform diagnosis and improve teaching processes and student learning. While there is a long tradition of learning outcomes' assessment within institutions' courses and programmes, emphasis on learning outcomes has become more important in recent years. Interest in developing comparative measures of learning outcomes has increased in response to a range of higher education trends, challenges and paradigm shifts.

AHELO aims to complement institution-based assessments by providing a direct evaluation of student learning outcomes at the global level and to enable institutions to benchmark the performance of their students against their peers as part of their improvement efforts. Given AHELO's global scope, it is essential that measures of learning outcomes are valid across diverse cultures and languages as well as different types of higher education institutions (HEIs).

The purpose of the feasibility study is to see whether it is practically and scientifically feasible to assess what students in higher education know and can do upon graduation within and across these diverse contexts. The feasibility study should demonstrate what is feasible and what could be feasible, what has worked well and what has not, as well as provide lessons and stimulate reflection on how learning outcomes might be most effectively measured in the future.

The outcomes of the feasibility study will be presented in the following ways:

- a **first volume** of the feasibility study Report focusing on the design and implementation processes which was published in December 2012;
- the present **second volume** on data analysis and national experiences;
- the feasibility study **Conference** which will take place in Paris on 11-12 March 2013; and
- a **third** and final volume to be published in April 2013 on further insights (and which will include the conference proceedings).

READERS' GUIDE

The chapter numbering follows from the first volume of the report. Therefore this second volume starts with Chapter 7 (volumes 1 to 6 having been published in the first volume).

Chapter 7 examines the issues of validity and reliability and provides insights on scientific feasibility from the AHELO feasibility study data.

Chapter 8 presents the experience of the feasibility study from the point of view of participating countries. Each country starts with an overview of main challenges, main achievements and main lessons learnt, in the format of a poster prepared for the Conference and elaborates in further details.

Chapter 9 was prepared by the Chair of the Technical Advisory Group and gives the conclusions and suggestions of this group on the feasibility study.

Note on terminology

The AHELO feasibility study involved the participation of 17 higher education systems. In most cases, participation was at the national level although a number of systems also participated in the feasibility study at the regional, provincial or state levels. This was the case for Abu Dhabi (United Arab Emirates), Belgium (Flanders), Canada (Ontario), and the United States (Connecticut, Missouri and Pennsylvania). For simplicity and ease of reading, all higher education systems are referred to as “countries” or “participating countries” in the report, irrespective of the national or sub-national level of participation.

Abbreviations used in the feasibility study report

AACC	American Association of Community Colleges
AAC&U	Association of American Colleges and Universities
AASCU	American Association of State Colleges and Universities
AAU	Association of American Universities
ACE	American Council on Education
ACER	Australian Council for Educational Research
AERA	American Educational Research Association
AHELO	Assessment of Higher Education Learning Outcomes
AMAC	Australian Medical Assessment Collaboration
AMK	Ammattikorkeakoulu – Finnish institution of higher education comparable to a university of applied sciences
APA	American Psychological Association
APEC	Asia-Pacific Economic Cooperation
ATAV	Adaptation, Translation And Verification
BA	Bachelor of Arts

BMD	Bachelor-Master-Doctorate (degree structure)
CAE	Council for Aid to Education
cApStAn	Linguistic Quality Control Agency
CHEPS	Centre for Higher Education Policy Studies
CLA	Collegiate Learning Assessment
CPR	Indiana University Center for Postsecondary Research
CRT	Constructed-Response Task
	<i>Within the AHELO feasibility study, different types of constructed-response items were used entailing different types of responses (short and extended responses, performance tasks, etc.). For simplicity within the Report, constructed response items take the abbreviation of a constructed-response task, or CRT.</i>
DIF	Differential Item Functioning
ECTS	European Credit Transfer and Accumulation System
EDPC	Education Policy Committee
EHEA	European Higher Education Area
EI	Education International
EQF	European Qualifications Framework
ETS	Educational Testing Service
EU	European Union
EUA	European University Association
EUGENE	European and Global Engineering Education academic network
FCI	Faculty Context Instrument
GDP	Gross Domestic Product
GNE	Group of National Experts
GRE	Graduate Record Examination
HEI	Higher Education Institution
IAU	International Association of Universities
IC	Institution Coordinator
ICC	Item Characteristic Curves
ICI	Institution Context Instrument
IDP Australia	International Development Programme
IEA	International Association for the Evaluation of Educational Achievement
IEA DPC	International Association for the Evaluation of Educational Achievement Data Processing and Research Center
IMHE	OECD Higher Education Programme (formerly Programme on Institutional Management in Higher Education)
IMHE GB	IMHE Governing Board
INES	OECD's Indicators of Education Systems (framework)
IRT	Item Response Theory
ISCED	International Standard Classification of Education
IRT	Item Response Theory
IUT	Institut Universitaire de Technologie
JSON	JavaScript Object Notation
LEAP	Liberal Education and America's Promise
LS	Lead Scorer
MA	Master of Arts
MAPP	Motivational Appraisal of Personal Potential
MCQ	Multiple Choice Question
MOOC	Massive Open Online Courses
MSC-AA	Medical Schools Council Assessment Alliance
NAEP	National Assessment of Educational Progress

NAICU	National Association of Independent Colleges and Universities
NASULGC	National Association of State Universities and Land-Grant Colleges
NC	National Centre
NCME	National Council on Measurement in Education
NIER	National Institute for Educational Policy Research
NILOA	National Institute for Learning Outcomes Assessment
NPM	National Project Manager
NSSE	National Survey of Student Engagement
OECD	Organisation for Economic Co-operation and Development
PIAAC	OECD Survey of Adult Skills (formerly Programme for International Assessment of Adult Competencies)
PISA	OECD Programme for International Student Assessment
PPP	Purchasing Power Parity
PWB	Programme of Work and Budget
QAA	Quality Assurance Agency for Higher Education
SCG	Stakeholders' Consultative Group
SCI	Student Context Instrument
STEM	Science, Technology Engineering and Mathematics
TA	Test Administrator
TAFE	Technical And Further Education
TAG	Technical Advisory Group
TALIS	OECD Teaching and Learning International Survey
TECA	Tertiary Engineering Capability Assessment
TRP	Technical Review Panel
UAT	User Acceptance Testing
UCTS	UMAP Credit Transfer Scheme
UIS	UNESCO Institute for Statistics
UMAP	University Mobility in Asia and the Pacific
UNDP	United Nations Development Program
UNESCO	United Nations Education Science and Culture Organization

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 evaluation of the scientific feasibility of AHELO's rests on the assessment of its capacity to produce valid and reliable results across different countries, languages, cultures and institutional settings. In the AHELO feasibility study context, scientific feasibility was defined as the “capacity of developing assessment instruments that would provide valid and reliable results across different countries, languages, cultures and institutional settings”.

This chapter presents an overview of the data collected and analyses conducted as part of the AHELO feasibility study in order to assess the scientific feasibility of the instruments that were used in the study¹. It focuses on the evidence collected to determine the validity and reliability of the instruments used for the feasibility study.

The analyses and results presented for the evaluation of the quality of the instruments used are to be interpreted in the “proof of concept” spirit of the feasibility study. The intent of the feasibility study was not to produce instruments to be re-used for a later study, but rather to prove the concept that such instruments can be developed and that they can produce valid and reliable results on an international scale, i.e. across different countries, languages, cultures and institutional settings.

Validity and reliability concepts for assessing scientific feasibility

A quality AHELO instrument needs to produce results that are both valid and reliable. Any instrument must produce valid results, i.e. results reflecting what it is intended to be measured, as well as reliable results, i.e. producing the same results over repeated measures. Both validity and reliability are essential criteria for any quality instruments.

In the context of the AHELO feasibility study, scientific feasibility depends on two questions:

- Are the instruments valid – i.e. do they measure what they are designed to measure, allowing for results and inferences to be considered valid?
- Are the instruments reliable – i.e. do they provide stable and consistent results over repeated measures allowing for results to be replicable?

A quality “international” instrument must also produce valid and reliable results across different countries, languages and cultures. An additional requirement for instruments administered in an international context is that they also need to provide evidence of validity and reliability across different countries, languages and cultures. Accordingly, supplementary analyses are required to verify that the instruments are valid and reliable not only within but also across countries, languages and cultures.

Validity

Validity is a broad concept that involves making appropriate interpretation and uses of test scores². It refers to the instrument's capacity to measure what is intended to be measured and to provide evidence that supports inferences about the characteristics of individuals being tested. It is often defined as the extent to which the instrument is doing the job intended and

that the actions undertaken on the basis of test scores are appropriate and supported by evidence (Van Essen, 2008). Strictly speaking, validity does not apply to the test itself, but rather to the inferences that can be made about the test results and how they are being used.

Validity requires that the purpose and inferences to be drawn from test scores be stated from the outset. It is essential that the purpose, the intended interpretation or inferences to be made, are clearly and explicitly stated right from the start, prior to the instrument development process as only a clear and well defined purpose can lead to an instrument development process best aligned with the intended inferences to be made (AERA, APA and NCME, 1999).

The evaluation of instrument validity requires the collection of a variety of evidence to support different types of validity. In the AHELO feasibility study, consistent with the AHELO Technical Standards (AHELO Consortium, 2012a) and the Standards for Educational and Psychological Testing (AERA, APA and NCME, 1999)³, four types of validity are considered as part of the instrument validation process:

- *Construct validity* refers to the extent to which all items are measuring the same construct⁴. For the AHELO feasibility study, evidence of construct validity is assessed through psychometric and statistical procedures used to analyse whether the instrument captures a single dimension of the underlying student ability⁵.
- *Content validity* refers to the extent to which the assessment instrument adequately represents the content and competencies of the domain of interest. For the AHELO feasibility study, evidence of content validity is assessed by analysing the extent to which the instrument development process⁶ ensured that the content was appropriate by:
 - using internationally agreed upon assessment frameworks to develop the instrument within appropriate domain contexts;
 - reaching international consensus on the mapping of items to a clear conceptual structuring of the domain defined in the framework;
 - ensuring that the instrument includes a broad and balanced set of test items that reflect the assessment framework; and
 - collecting student feedback during focus groups and cognitive labs.
- *Face validity* refers to the extent to which the instrument is perceived as valid by stakeholders at face value. For the AHELO feasibility study, face validity is evaluated by considering:
 - feedback collected during the instrument development process from Expert Groups, the TAG, HEIs, and stakeholders (see Volume 1); and
 - students' reaction to the instrument during the field implementation, i.e. student engagement and reported effort in responding to the test, as well as student perceived relevance of the test.

- *Concurrent validity* refers to the extent to which test results on students' learning outcomes vary with related measures of student abilities. This is assessed through correlations between the test scores and other proxies of student abilities such as reported academic performance.

Reliability

Reliability means that test results are consistent and stable across different testing situations.

Reliability is the second condition that an instrument must satisfy, in addition to validity. Reliability does not imply validity, which must be independently established.

An instrument's degree of reliability can be affected by a number of different factors. Test scores are considered to be the result of two components: the true ability level tested and random factors that may affect the student performance on the test. Such factors include the number and the quality of items, the conditions under which the instrument is administered, the characteristics of the students such as the effort they put into the test, and the quality and consistency of scoring for the constructed-response tasks.

Stable results suggest that the observed student scores are more likely to reflect true scores.

The more consistent or stable the results are, the more one can be confident that the results represent the students' "true" scores, and that measurement errors are minimised.

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. The ratio is represented as a reliability index ranging from 0 to 1, where 0 represents very poor consistency and 1 perfect consistency. As per the AHELO Technical Standards, item-level reliability indexes of 0.80 or higher are regarded as acceptable, indicating reasonably small measurement error.

Evidence on scientific feasibility collected during the AHELO feasibility study

Some individual test items are dysfunctional and must be removed before validity and reliability can be assessed. The data collected during the AHELO feasibility study field work provide the evidence base for evaluating a range of validity and reliability dimensions. Prior to undertaking these analyses, it is however important to review the quality and functioning of individual test items to remove those that are dysfunctional.

Review of item quality and functioning

It is necessary to conduct a quality check to review whether the study produced items that function well for the target population. Before data analyses can be run, items that do not function as expected, or that function differently for different sub-populations without explanations for these variations, must be removed from the data sets. It is common and expected in large-scale surveys for some items to be removed – even more so in the case of the AHELO feasibility study since there was no field trial to test the psychometric performance of various items and revise the instruments on this basis prior to the main administration.

Removal of non-functioning items

Items not meeting psychometric standards⁷ are deleted from final analyses. The item analyses conducted for the AHELO feasibility study consists in examining standard items statistics including item difficulty, item discrimination, item-to-test correlations, item-to-total correlations, and count and percentage of each response category. Other analyses are also conducted as the test scores were analysed with the Rasch item response model: goodness of fit, item characteristic curves (ICCs), item-by-country interaction and differential item functioning (DIF). Each of these item statistics provides an indication of the item quality. For example, an item with low discrimination indicates that the item does not discriminate well between students with higher and lower levels of ability. Such information at the item level, along with indications from other item statistics, is used as a basis for item deletion.

Non-functioning items can be removed on a country basis. The psychometric qualities of items are assessed for the entire population participating in each strand, but also at the country level. This additional level of analysis allows for the identification of items that do not function well only for some countries. While removing non-functioning items for specific countries makes the all-countries instrument a more valid and reliable measure of the construct, it is important to ensure that the items removed do not affect the framework coverage in a particular country. It is also worth noting that removing items by country may include items where students performed both well and poorly in that country.

The small number of items removed from the generic skills instrument indicates an instrument with good overall item quality. Of the generic skills instrument's 55 multiple-choice items, three items are removed for all countries, and up to three are removed for three individual countries⁸ (Table 7.1). The student scores distribution indicates that instrument was well targeted to students. Item difficulty indices indicate that only one item was answered correctly by more than 70% of the students, while 14 out of 55 items are failed by 30% or more students. The overall item quality of the generic skills instrument is particularly good given that pre-validation of the instrument consisted in qualitative cognitive labs and did not include quantitative validation with focus groups.

Table 7.1 - Number of items removed by country - Generic Skills

Item	Countries			
	All countries	Country 1	Country 2	Country 8
MCQ6		1	1	
MCQ12	1			
MCQ15				1
MCQ17	1			
MCQ18				1
MCQ36	1			
MCQ51				1
Total	3	1	1	3

The relatively small number of items deleted from the economics instrument indicates an instrument with sufficient overall item quality. Of the economics instrument's 61 items, one multiple-choice item is removed for all countries, while up to 13 multiple-choice items are removed for one country (Table 7.2). Item difficulty indices indicate that only one item was answered correctly by more than 70% of the students and 28 items were answered correctly by less than 30% of the students. In terms of item discrimination, results indicate that at the international level, 19 items are problematic with an item-to-test correlation below 0.20. The overall item quality of the economics instrument is fairly good given that while focus groups were conducted during the development phase of the instrument, no field trial allowed for refining the items based on their performance with larger groups of students.

Table 7.2 - Number of items removed by country - Economics

Item	Countries						
	All countries	Country 1	Country 2	Country 3	Country 4	Country 5	Country 6
MCQ3						1	
MCQ5				1		1	
MCQ6					1	1	
MCQ9				1	1		
MCQ15						1	
MCQ16	1						
MCQ18				1			
MCQ19				1		1	1
MCQ22			1	1	1	1	
MCQ23			1			1	
MCQ24			1				1
MCQ25						1	
MCQ26			1				
MCQ28				1			
MCQ29			1	1	1	1	
MCQ30			1			1	
MCQ31						1	
MCQ34						1	
MCQ36		1					
MCQ38			1			1	
MCQ39				1			
MCQ41					1		
CRT2F			1				
Total	1	1	8	8	5	13	2

The relatively small number of items deleted from the engineering instrument indicates an instrument with sufficient overall item quality. No item is deleted for all countries from the engineering instrument. The review of item statistics at the country level indicate however that certain multiple-choice items and constructed-response tasks did not function as expected for some countries. Of the 30 engineering multiple-choice items and 3 constructed-response tasks

(with 21 sub-items), between two and eleven items are removed for different countries (Table 7.3).

Table 7.3 - Number of items removed by country - Engineering

Item	Country 1	Country 2	Country 3	Country 4	Country 5	Country 6	Country 7	Country 8	Country 9
MCQ2				1		1			
MCQ4		1							
MCQ5	1			1	1	1			
MCQ6									
MCQ8							1		
MCQ10				1		1			
MCQ12					1				
MCQ14				1		1			
MCQ15	1								1
MCQ16			1						
MCQ18								1	
MCQ19				1	1			1	
MCQ20					1		1		
MCQ21					1				
MCQ23				1					
MCQ24						1			1
MCQ25				1		1			
MCQ26		1		1			1	1	
MCQ27	1								
MCQ28				1				1	
MCQ30				1	1				
CRTM13		1	1					1	
CRTM32				1	1				
Total	3	3	2	11	7	6	3	5	2

Differential item functioning (DIF) analyses

In an international study, it is critical to ensure that items have a similar level of difficulty across the different countries used to make comparison in student performance. Consistency of item difficulty across countries and languages is of particular importance. An item measuring the same underlying construct across countries and languages should have a similar difficulty level relative to other items across countries and languages. Item statistics indicating that an item was poorly achieved in one country may indicate that the item did not function as expected due to unintended features of the item such as errors introduced in translation, or the use of content unknown of the student population in that country. The same item difficulty consistency is expected across countries of institutions sharing similar characteristics.

Differential item functioning (DIF) analyses are conducted to further understand differences in performance of different student sub-populations. Cross-cultural surveys such as the AHELO feasibility study require additional specific analyses to evaluate the extent to which it is possible to validly and reliably assess student performance within different institutional,

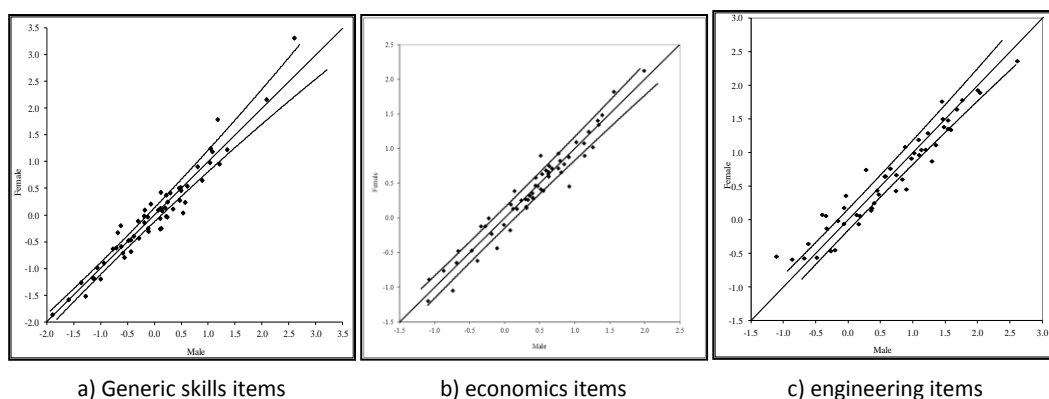
linguistic and cultural contexts. The purpose of these analyses is to further understand differences in performance of different student sub-populations and to verify that the items used did not introduce biases for specific sub-groups of the target populations. In the AHELO feasibility study, differences in student performance are examined between genders, across countries of institutions, countries, and test languages.

Item functioning across genders

Most items show no significant differences in performance between genders. In all three strands, results of gender DIF analyses indicate that most items have similar difficulty for both males and females (Figure 7.1). Some items however, show relatively large differences in student performance when comparing male and female students. Student performance results for the engineering instrument indicate for example that some items were better achieved by females, while others were better achieved by males.

Further analyses would be needed to identify the underlying reasons for these differences by gender. Closer look at those items, with test developers' insights, would be required to better understand the differences in performance between males and females, identifying possible item features that would benefit one country over the other.

Figure 7.1 - Gender DIF analyses⁹



Item functioning across institution types

Three institutional characteristics are used as a basis for comparison across the different types of higher education institutions. In the absence of an international classification of higher education institutions besides the Carnegie classification in the USA and the U-map initiative in the EU context, three institutional characteristics are selected as a basis for comparison amongst different higher education institution profiles: i) the size of the institution, ii) the highest degree offered at the institution and iii) the institution emphasis on research and teaching.

Little difference in student performance is observed when using the institution size (small/medium/large) as a basis for comparison. In all three strands, being a small, medium or

large size institution has little impact on item difficulty, with two noticeable exceptions: large differences in item difficulty are observed when comparing small to medium size institutions for the economics instrument, and large to medium size institutions for the engineering instrument. With the economics instrument, many items are more difficult for students in medium size institutions as opposed to students in large size institutions. For the engineering instrument, many items are more difficult for students in medium size institutions as opposed to students in large size institutions.

Figure 7.2 - Institution DIF analyses: Size (small versus medium)

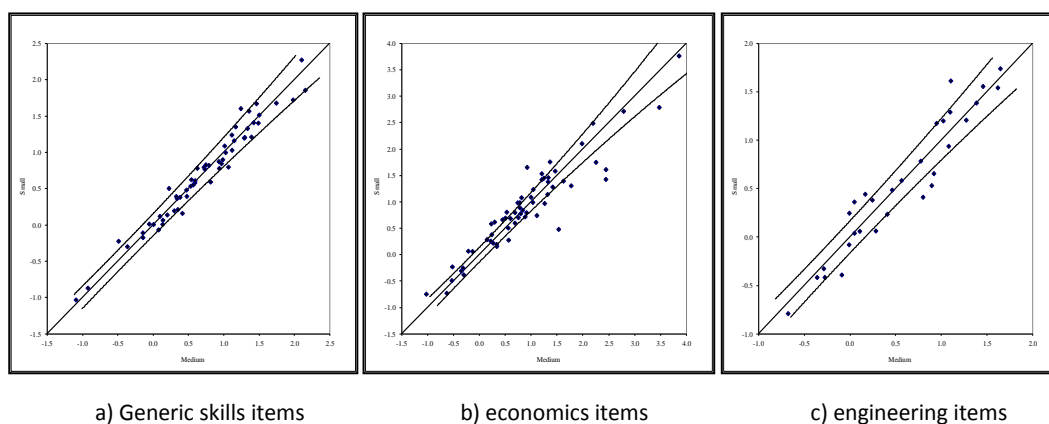
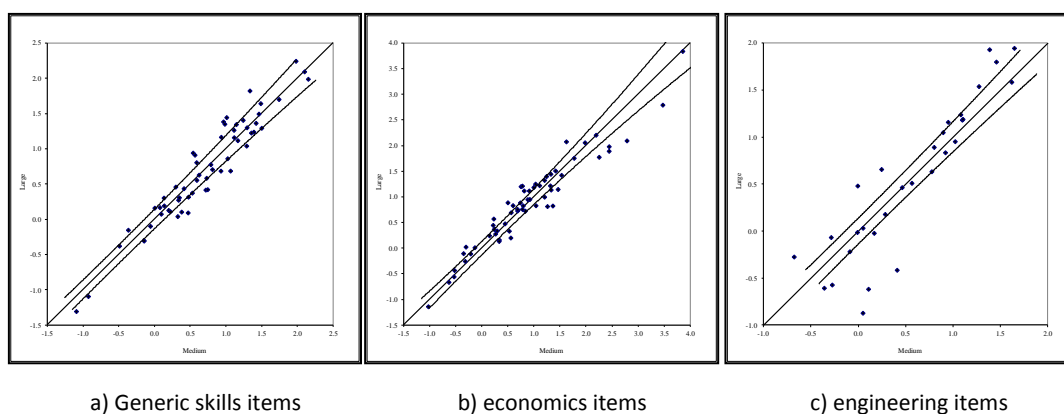


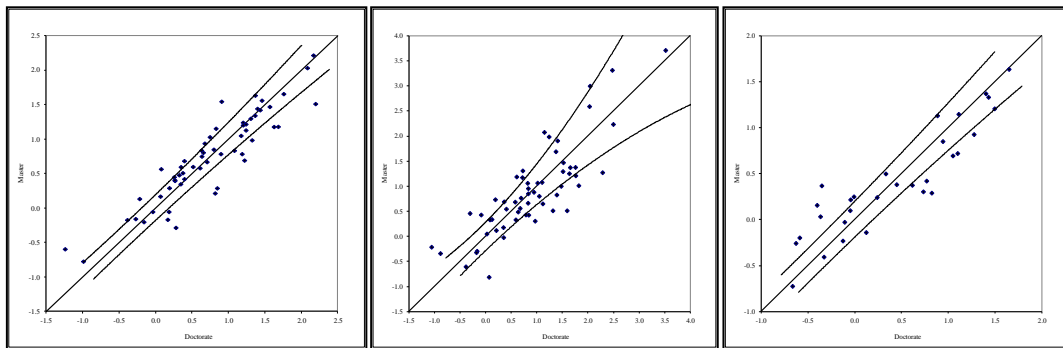
Figure 7.3 - Institution DIF analyses: Size (medium versus large)



Differences in student performance are observed when using the institution highest degree offered (baccalaureate, master and doctorate) as a basis for comparison. In all three strands, the item difficulty varies across institutions with different degree granting profiles, with some items being easier for students attending baccalaureate degree granting institutions, or master

degree granting institutions, and some items being easier for students attending doctorate degree granting institutions.

Figure 7.4 - Institution DIF analyses: Highest degree offered (doctorate versus masters)

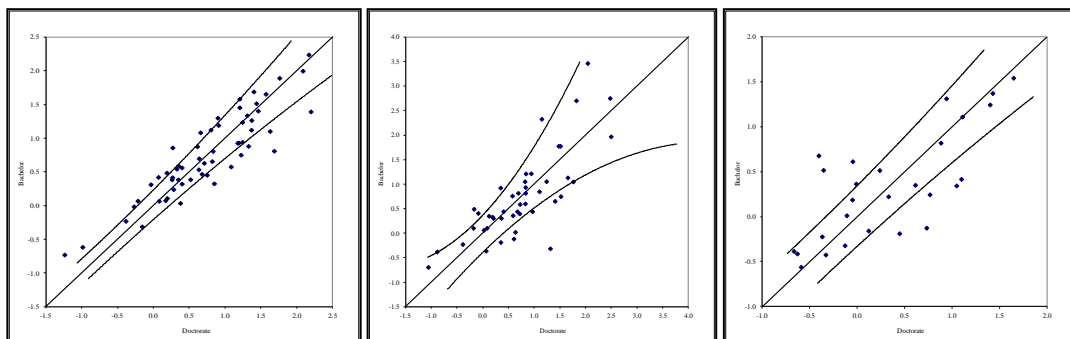


a) Generic skills items

b) economics items

c) engineering items

Figure 7.5 - Institution DIF analyses: Highest degree offered (doctorate versus baccalaureat)

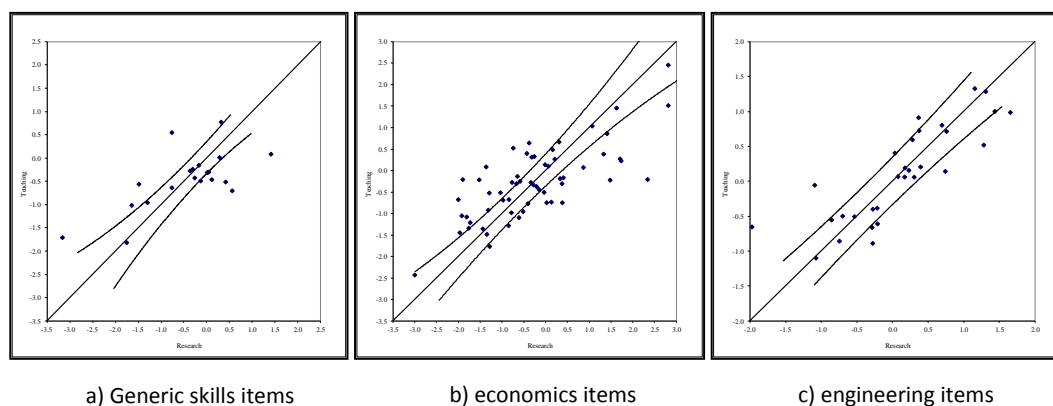
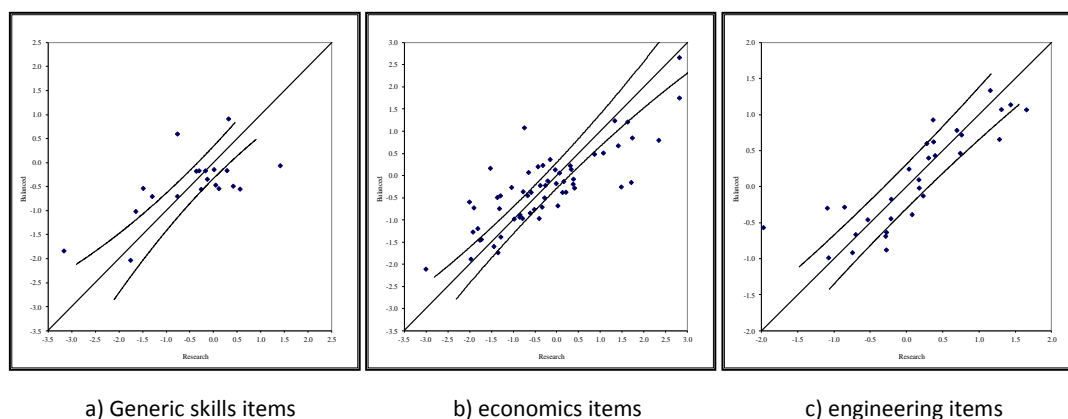


a) Generic skills items

b) economics items

c) engineering items

Differences in student performance are observed when using the institution emphasis on research and teaching (research, teaching, and research/teaching balance) as a basis for comparison. In all three strands, the item difficulty varies across institutions with different teaching/research emphasis, with some items being easier for students attending institutions where the emphasis is on teaching, or balanced between teaching and research, while some items were easier for students attending institutions where the emphasis is on research.

Figure 7.6 - Institution DIF analyses: Emphasis (research versus teaching)**Figure 7.7 - Institution DIF analyses: Emphasis (research versus research/teaching balance)****Results indicate difference in student performance across the different institution types.**

Although not much difference in student performance is observed when comparing institutions based on their size, large differences are observed when comparing them using the highest degree offered and their emphasis on research and teaching. However, with results portraying a variety of patterns for the different types of institutions in all three strands, no conclusive insights can be made. Additional analysis would be required to further explore the reasons as to why those items performed differently for different types of institutions.

Item functioning across countries

Many items do not function as expected for some countries¹⁰. In all strands, results of DIF analyses comparing student performance across countries indicate that many items do not function as expected for some countries. Some items are more difficult than expected for some countries while others are easier. Furthermore, results indicate that multiple-choice items

provide more consistent results than constructed-response tasks. When comparing student performance in one country to all other countries at the item level, differences observed are more significant for the constructed-response tasks in the economics and engineering strands. The student performance differences across countries are less for the generic skills constructed-response tasks, suggesting greater consistency of item difficulty levels across countries.

Table 7.4 - Item functioning differences across countries

		Items easier than expected in at least one country		Items harder than expected in at least one country	
		n	%	n	%
Generic Skills (9 countries)	CRT	3	50.0	2	33.3
	MCQ	19	36.5	16	30.8
Economics (6 countries ¹¹)	CRT	12	92.3	9	69.2
	MCQ	25	53.2	12	25.5
Engineering (9 countries)	CRT	14	66.7	13	61.9
	MCQ	16	61.9	7	23.3

Further analyses are needed to identify the underlying reasons for these country differences.

A closer look at those items performing differently for different countries is required to better understand the differences in student performance across countries, identifying possible item features that would benefit one country over another.

Item functioning across languages

Many items do not function as expected for some languages¹². In all strands, results of DIF analyses comparing student performance based on the test language indicate that many items did not function as expected for some languages. Some items are more difficult than expected for some languages while other items are easier.

Table 7.5 - Item functioning differences across languages¹³

		Items easier than expected in at least one language		Items harder than expected in at least one language	
		n	%	n	%
Generic Skills (7 languages)	CRT	4	66.7	6	100
	MCQ	29	29.5	35	67.3
Economics (7 languages ¹⁴)	CRT	7	53.8	9	69.2
	MCQ	34	72.3	23	48.9
Engineering (7 languages)	CRT	19	63.3	18	60.0
	MCQ	12	57.1	12	57.1

Constructed-response tasks in the generic skills strand show significant differences in student performance across languages. When student performance in one language is compared to student performance in all other languages combined, items show more significant differences for the constructed-response tasks in the generic skills strand. This pattern of student performance differences across languages between the two item types is less clear for the economics and engineering instruments.

Further analyses are needed to identify the underlying reasons for student performance differences in the different languages. A closer look at the items may reveal language errors or biases resulting from the translation and adaptation process. The fact that constructed-response tasks utilise generally more written materials in comparison to multiple-choice items may explain why there is a bigger difference in performance when comparing languages are observed for this type of item. Out of the three assessment instruments, the generic skills instrument is the one using the most written material for the construct-response tasks.

Validity evidence

Different types of evidence are collected throughout the feasibility study to determine the validity of instruments¹⁵ used. The validation process used to substantiate result interpretations considers four types of validity evidence: construct validity, content validity, face validity and concurrent validity, using both qualitative and quantitative evidence collected throughout the feasibility study, from the instrument development process to final data analysis.

Construct validity

The three assessment instruments display reasonable levels of construct validity evidence. Exploratory and confirmatory factor analyses were conducted to assess the dimensionality of each instrument. In addition, results of analyses of fit statistics¹⁶ indicate that constructed-response tasks and multiple-choice items scale well onto a single dimension for all three assessment instruments, thereby indicating that each instrument is measuring one common construct¹⁷.

Results indicate that the overall scale could be divided into complementary sub-scales. The psychometric analysis results also indicate that for each instrument, items could also be divided up into different subsidiary complementary sub-scales. For example, the overall scale could be broken down into sub-scales representing the five economics learning outcomes stated in the framework. However, given that the purpose of the feasibility study was not to develop comprehensive instruments, the number of items across these learning outcomes is insufficient for a detailed analysis by such sub-scales. This underlines the scope for further improving the construct validity of all three instruments.

Content validity

Expert consensus

Evidence of content validity of the generic skills instrument is not fully demonstrated through expert consensus. The instrument development process for this strand relied on the adaptation and translation of an existing generic skills test for international use. The development of an international framework was only added subsequently since it was not included in the original study design. Time constraints limited the scope for seeking expert consensus on both the framework and the instrument. As a result, further expert consultation is needed to document content validity evidence for this instrument.

Evidence of content validity of the economics and engineering instruments is provided through expert consensus. The instrument and framework development process included expert consensus on both the frameworks and the assessment instruments. In this respect, the instrument development process adopted ensured that the two discipline instruments' content is appropriate in relation to their respective framework.

Student feedback

Feedback from the generic skills cognitive labs showed that the constructed-response tasks were attractive to students. Collected during the instrument development process¹⁸, students' feedback indicates that in general students found the constructed-response tasks engaging and interesting. Many students commented on the fact that they were not familiar with this type of tasks. However, many also stated that the documents provided contained a great deal of information and that, combined with the questions, forced them to think more deeply about the test material.

Students also reacted positively to the draft economics and engineering constructed-response tasks. Students' feedback collected during the focus groups¹⁹ indicates that more than half of the students agreed, or strongly agreed, that the constructed-response tasks covered topics relevant to their programme. In general, students found the tasks to be challenging, interesting, clear and comprehensive. Their comments also indicated that they felt the tasks were very much related to the real world and was a "good tool to assess our knowledge" (economics strand) and "an efficient tool to measure a broad range of knowledge" (engineering strand). The most common complaint was that the time available to do each task was too short.

Further content validity evidence for the two discipline instruments is still required to fully confirm content validity. Countries and systems which joined the study at a late stage could not be part of the discipline expert groups and were thus not much involved in framework and item review²⁰. Feedback from participating countries on the instrument development process revealed a need from participants to be more deeply involved in instrument development, with several National Project Managers being somewhat critical of the consultation and revision processes and reporting that very fundamental comments on some of the items were “communicated without much reaction, or with insufficient will from test developers to apply the suggested changes” (Brese and Daniel, 2012). This suggests the need to collect further content validity evidence for the two discipline instruments.

Face validity

Face validity is assessed through several indicators. For the AHELO feasibility study, in addition to evidence of face validity collected during the instrument development process from Expert groups, the TAG, HEIs, and stakeholders (see Volume 1), evidence is also collected from students’ reactions to the instrument during field implementation. Three indicators are used: student engagement with the assessment, student reported effort put into the assessment and student perceived relevance of the test.

Student engagement with the assessment

Students spent a good deal of time responding to the AHELO assessments. In all three strands, the time students spent responding to the tests suggest a good degree of student engagement. For the generic skills strand, students spent, out of the 90 minutes allocated to the constructed-response tasks section, an average of 56 minutes, and students spent all of the set 30 minutes on the multiple-choice items section. For the economics strand, out of the total 90 minute assessment, students spent an average of 75 minutes responding to the test. For the engineering strand, students spent an average of 65 minutes out of the 90 minutes allocated for the entire test.

The low levels of non-response indicate good levels of student engagement with the instruments. In all three strands, item-level non-response remains sufficiently low to support claims of face validity. For the generic skills instrument, student engagement in terms of non-response shows that students took the assessment instrument seriously with non-response rates in the order of 12 to 13% for the multiple-choice items²¹ (Table C1). For the economics instrument, levels of non-response are slightly higher, in the order of 22% for multiple-choice items, and 7 to 17% for the constructed-response tasks (Table C2). For the engineering instrument, levels of non-response are in the order of 12 to 16% for multiple-choice items and 2-3% for constructed-response tasks (Table C3). These low levels of non-response indicate that students took the assessment instruments seriously. The highest levels of non-response for the economics instrument require further consultation and investigation to identify the underlying reasons which led students to skip specific items.

Student reported effort

Students reported putting a good deal of effort into the AHELO assessments. The students' reported effort provides an indication of individual perceptions of the assessments. This information was collected via the student questionnaire asking students how much effort they put into the test on a 4-point scale²². At the international level, self-reported effort put into the assessment is about 2.75 for the generic skills strand (Figure C1) and 2.5 for the economics and engineering strands (Figures C2 and C3). Self-reported effort across countries suggests fairly similar patterns in all strands and countries, with reasonably high levels of student effort put into the tests and limited variations across countries. This equates well with the above results on student engagement, providing some indication that the students who decided to show up and take the test made a reasonable effort to respond to the assessment instruments.

Self-reported effort by field of education for students participating in the generic skills strand also reveals limited variations across fields. Self-reported effort results indicate some small differences for certain fields, with students in education, social sciences, business and law putting slightly more effort into the test while students in services report lower levels of effort (Figure C4). However, the differences in self-reported effort are much lower across fields of education than they are across countries (Figure C1).

Students' perceptions of the relevance of the tests

Students' perceptions of the educational and professional relevance of the instruments vary across strands. As part of the student questionnaire, students were asked to rate, on a 5-point scale²³, how relevant the test materials were to their current degree and to future professional practice. While direct comparisons of perceptions of relevance across strands cannot be stretched too far given substantive and cultural differences, results indicate students' perceptions vary across strands. Students participating in the generic skills strand saw greater relevance of the instrument in relation to their future profession rather than current study, not surprisingly given the generic focus of the assessment. By contrast, students participating in the two discipline strands saw greater educational than professional relevance, not surprisingly given the disciplinary focus of these assessments (Figure C5).

Students' perceived relevance also reveals some differences across fields of education. Results indicate some differences across fields of education with social science, business, law and engineering students seeing the test as more relevant to their current study and future profession than students in the humanities, arts, health, welfare and service fields (Figure C6).

Concurrent validity

Two indicators are used as criteria to provide concurrent validity evidence. In the context of the AHELO feasibility study, the two indicators that serve as criteria for concurrent validity are self-reported academic performance and student satisfaction with their educational experience.

AHELO test scores and self-reported academic performance

Results show a correlation between students' AHELO test scores and their self-reported academic performance only for the engineering strand. In the generic skills and economics strands, there appears to be only a very mild relationship between these reports and test scores. In the engineering strand, there seems to be a stronger relationship between academic performance and test scores. Students in this strand reporting below average academic performance had an AHELO test score lower than those reporting being above average.

The strength of the relationship between AHELO scores and self-reported academic performance varies across countries. Analysis conducted at the country level shows, in all three strands, that the relationship between the two indicators varies in terms of direction and strength, suggesting differing levels of concurrent validity depending on national contexts (Figures C7, C8 and C9). However, those correlations should be interpreted with caution as they rely on a self-reported indicator and require further exploration of the influence of cultural factors, the impact of population and sampling and interactions with curriculum and pedagogy.

AHELO test scores and student satisfaction²⁴

The strength of the relationship between AHELO scores and students' overall education satisfaction varies across countries. Analysis conducted at the country level showed in all three strands that the relationship between the two indicators varies in terms of direction and strength. In the generic skills and economics strands, the relationship between the students' AHELO test scores and their overall education experience satisfaction show inconsistent patterns (Figures C10 and C11). In the engineering strand, there seems to be a low relationship between students' satisfaction with education and their test scores, although it is not prevalent in all countries (Figure C12). Further analysis is required to better take into account institution-level variability to see if student satisfaction can be used as a good criterion for concurrent validity.

Reliability evidence

Reliability indices

Overall reliability indices

The feasibility study produced instruments with "acceptable" to "good" levels of reliability. For the generic skills and economics instruments, the final reliability is respectively 0.83 and 0.84 when using final plausible values²⁵. In both cases, reliability estimates are above 0.80, the threshold set in the AHELO Technical Standards. For the engineering instrument, the final reliability is 0.75 when using final plausible values, which is lower than the threshold specified in the AHELO Technical Standards, but still within scope for an assessment instrument of this kind (Table C6).

Examination of reliability indices at the country level shows less reliable results for some of them. Overall, reliability indices remain acceptable for about half of countries participating in the generic skills, and half of the countries participating in the economics, i.e. with reliability

indices of 0.70 or higher. For countries participating in the engineering strand, the highest reliability level achieved was 0.66 for one country (Table C6).

Reliability indices using institution-level aggregated data

Reliability analyses using data aggregated at the institutional level suggest “acceptable” to “good” levels of reliability in all three strands. The lowest reporting level for the AHELO feasibility study is the institution. Therefore, it makes sense to estimate reliability using data aggregated at the institutional level²⁶. Aggregate reliability for the generic skills strand reached a mean of 0.72, while it reached 0.95 for the economics strand and 0.88 for the engineering strand.

Examination of reliability indices using data aggregated at the institutional level indicates less reliable results for some countries. The aggregate reliability for the generic skills strand ranged from 0.62 to 0.83 between countries. For the economics strand, it varied from 0.53 to 0.99 and from 0.45 to 0.66 in the engineering strand.

Constructed-response tasks scoring reliability

Inter-scorer reliability statistics²⁷ can be considered “fair” to “good” in all three strands. Inter-scorer reliability statistics vary from across the three strands (Tables C7 to C9). Direct comparisons of statistics amongst strands should be made carefully given that no correction for differences in the range of scoring points has been made. Wider ranges of scoring points, e.g. from 1 to 6 points versus from 0 to 1 or 2 points, will lead to larger differences in score points.

Scoring of the generic skills constructed-response tasks meets the standards. Summary inter-scorer reliability statistics for the generic skills constructed-response tasks indicate very little divergence amongst the three scoring criteria or across the two constructed-response tasks²⁸. The mean difference statistics indicate that around half of the scorers gave the same score-point and around a third of allocated scores were within one score-point difference. The intraclass correlations of each constructed-response tasks approach the threshold standards of 0.85, while kappa is lower, indicating a moderate effect (Table C7).

Scoring of the economics constructed-response tasks also meets the standards. Summary inter-scorer reliability statistics for the economics constructed-response tasks indicate that percentage agreement statistics are higher than for the generic skills instrument, likely related to the smaller range in the score categories for items in the economics instrument²⁹ (Table C8). This pattern is also observed between CRT1, displaying around 90% agreement, and CRT2 where percentage agreement is around 80% likely related to the greater range in the score categories for CRT2 items.

Scoring of the engineering constructed-response tasks also meets the standards. Summary inter-scorer reliability statistics for the engineering constructed-response tasks indicate that the percentage absolute agreement statistic sits at around 80% for all constructed-response tasks which can be considered fair to good³⁰ (Table C9). This places the engineering scoring reliability between that of the economics instrument (highest reliability) and the generic skills instrument (lowest reliability), although it would be necessary to correct the generic skills statistics for the range of score categories.

Cross-country inter-scorer reliability results indicate that reliable results can be obtained. In addition to inter-scorer reliability analyses conducted within countries, two evaluation of cross-country inter-scorer reliability were conducted once scoring was completed, one in the generic skills strand and one in the engineering strand scoring. In both cases, results tend to indicate that it is feasible to score constructed-responses in a reliable way across countries, given that conditions such as the scoring rubric clarity, appropriate training and ongoing monitoring of reliability, are put in place. However, results from those two studies rely on limited number of student papers and small number of scorers, and should therefore be considered with caution.

Scoring of student responses may vary across countries but their rank ordering is very consistent. The approach chosen to explore cross-country inter-scorer reliability for the generic skills strand consisted in selecting two sets of student responses, one set originally written in English and translated in other languages, and one set in other languages translated in English. Translated student responses were then scored by Lead Scorers from five participating countries. Results indicate that there is a high degree of agreement within and across countries in scorers' judgments about task difficulty and relative answer quality. The score assigned to a response may vary somewhat across countries, but the scores are very consistent in the rank orderings of the quality of the responses across countries.

Scoring of student responses is consistent across countries when considering the tasks total scores. The approach chosen to explore cross-country inter-scorer reliability for the engineering strand consisted in selecting a set of student responses from countries that had students writing the test in English, and have those responses rescored by six Lead Scorers who had been involved in the scoring activities in their countries. Student responses were collected for each of the three constructed-response tasks and their 6 sub-items. Results indicate that it is possible to score constructed-response tasks across countries in a reliable way, as long as the scoring rubrics are unambiguous. The data indicates that constructed-response tasks with sub-items may lead to agreement in sum scores while indicating less reliable scoring when considering sub-items individually.

Other indicators on the quality of the test

Relation of item difficulty to student ability level

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. For the generic skills instrument, the variable map³¹ shows that in general the multiple-choice items are well spread across the student ability distributions, with some items slightly at a difficulty level above the average student's ability than is ideal (Figure C13). The distribution of the student abilities shows that the instrument is able to discriminate student abilities over a good range.

The distribution of student performance shows that the economics and engineering tests were too difficult. For the economics and engineering instruments, the variable maps indicate that students found the test relatively difficult, with more items sitting above the average student performance than below (Figures C14 and C15). The two variable maps show however that in general items were well spread with respect to the ability distribution.

The large proportion of “zero” scores for the economics and engineering constructed-response tasks³² also indicates that the items were too challenging for students. For the economics constructed-response tasks, the proportion of zero scores ranged from 35 to 90 percent depending on tasks and countries, while the range was between 20 and 70 percent for the engineering constructed-response tasks [Figures C16 and C17]. The large proportion of zero scores means that the instruments did not allow for testing a sufficiently large continuum of ability. Consequently, test results variation is limited by the fact that students were not provided with the opportunity to show their ability, given that the items were too difficult.

Item sensitivity to effort

Effort seems to have a greater impact on constructed-response tasks than on multiple-choice items. Students taking the test may interact with the two types of items in different ways. Preliminary analyses indicate the effort students put into a constructed-response task response has a sizeable impact on their scores. Comparison of the percentage of variance in student scores explained by student effort for each of the two item types indicates a greater impact of effort on scores for the constructed-response tasks in the generic skills and engineering strands (Figures C18, C19 and C20). More detailed analysis is needed to explore potential reasons for this kind of method effect on student performance.

Conclusions

Overall item quality and functioning

The AHELO feasibility study produced many items that functioned well. The relatively small numbers of mal-functioning items removed from analysis is an indication of the overall good quality of the instruments used for the feasibility study. However, when considering item performance for different sub-groups based on gender, institution types, countries and languages, results indicate significant differential item functioning. The constructed-response tasks were, for example, more sensitive to differential item functioning. Although differential item functioning does not necessarily mean item bias, further analyses would be required to understand the underlying reasons for the different patterns.

Overall assessment of validity

All three instruments have achieved reasonable levels of construct validity. Factor analyses and item fit statistics suggest that each of the three instruments is measuring one common construct. Although the evidence collected to support construct validity of all three instruments is sufficient for the purpose of the study, additional evidence would be required to fully support construct validity in a full-scale study.

The evidence collected also suggests that the instruments have achieved reasonable levels of content validity in the disciplinary strands. For the generic skills strand, the instrument development approach adopted did not involve a sufficient consultative process to reach an international agreement on the appropriateness of the instrument content. The instrument would require further consultation to provide evidence of content validity. For the two discipline-based instruments, agreement was reached by participating stakeholders and

experts on frameworks and selected learning outcomes, providing evidence of content validity for each of these two instruments. Additional evidence of content validity includes feedback collected from students during the instrument development phase, for all three instruments.

The evidence collected also suggests that the instruments have achieved reasonable levels of face validity in all three strands. In addition to evidence of face validity collected during the instrument development process from Expert Groups, the TAG, HEIs, and stakeholders (see Volume 1), evidence is also collected from students' reactions to the instrument during field implementation. Student engagement with the assessment in terms of time spent and test completeness, along with their self-reported effort put into the assessment indicates strong evidence of face validity for all three instruments. Results also indicate that when considering students' perceptions of educational and professional relevance of the instruments, students participating in the generic skills strand saw greater professional relevance of the instrument, while students participating in the two discipline-based strands saw greater educational relevance of the instruments.

Evidence on concurrent validity is less conclusive. Results indicate no strong correlations between self-reported academic performance, or student satisfaction, and AHELO scores, with the possible exception of the engineering strand. The evidence collected indicates that while the engineering instrument displayed some correlation, patterns are less clear for the generic skills and economics instruments. Furthermore, the relationships between the AHELO scores and self-reported academic performance or overall education satisfaction vary significantly across countries. Additional analyses with more direct measures of student abilities would therefore be needed to provide further evidence of concurrent validity.

Overall assessment of reliability

The three instruments provided reliable results. Results of reliability analyses provide evidence that the three instruments functioned reliably overall. However, when looking at reliability estimates at the institution or country levels, some low estimates suggest that improvements are needed to provide sufficient reliable results. One example of possible improvement is a better match of the test difficulty with students' abilities during the test development phase, eliminating the items that are deemed too difficult for students to respond.

Inter-scorer reliability can be considered "fair" to "good" in all three strands. Scoring in all three strands provided inter-scorer reliability indices meeting reliability standards. Similarly, results of two cross-country inter-scorer reliability studies also tend to indicate that it is feasible to score constructed-responses in a reliable way across countries. Although variations in scoring across countries were observed, scores were consistent in the rank orderings of the quality of the responses, which indicates that with appropriate training and monitoring, scoring of student responses can be done reliably across countries.

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. Although overall results provide sufficient evidence of the feasibility of obtaining valid

and reliable results, further evidence in an AHELO main study would be required to better understand and explain differences of instrument quality across different types of institutions, countries and languages.

REFERENCES

- AHELO Consortium (2012), *AHELO Technical Standards*,
[http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/imhe/ahelo/gne\(2012\)16&doclanguage=en](http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/imhe/ahelo/gne(2012)16&doclanguage=en)
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999), *Standards for educational and psychological testing*. American Psychological Association, Washington, DC.
- Brese, F. and T. Daniel (2012), *OECD Assessment of Higher Education Learning Outcomes (AHELO) Feasibility Study: Report on Quality Adherence*, IEA Data Processing and Research Center, Hamburg.
- Messick, S. (1989), Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed). (pp.13-103). American Council on Education. Macmillan, Washington, DC.
- Mislevy, R.J. (1991), "Randomization-based Inference About Latent Variables from Complex Samples", *Psychometrika*, No. 56, pp. 177-196.
- Mislevy, R.J., Beaton, A., Kaplan, B.A. & Sheehan, K. (1992), "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses", *Journal of Educational Measurement*, No. 29 (2), Wiley, pp. 133-161.
- OECD (2012), *Assessment of Higher Education Learning Outcomes Feasibility Study Report - Volume 1 - Design and Implementation*, Paris
<http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume1.pdf>
- Van Essen T. (2008), "Validity and Reliability – Considerations for the OECD Assessment of Higher Education Learning Outcomes", presented to the AHELO Country of National Experts, OECD Paris, 17-18 December 2008,
[http://www.oecd.org/officialdocuments/displaydocumentpdf?cote=EDU/IMHE/AHELO/GNE\(2008\)6&doclanguage=en](http://www.oecd.org/officialdocuments/displaydocumentpdf?cote=EDU/IMHE/AHELO/GNE(2008)6&doclanguage=en)

NOTES

- ¹ In presenting results of psychometric analyses, country names are not disclosed since groups of participating higher education institutions (HEIs) within each country/system are not representative of the higher education system in which they belong, and the purpose of the AHELO feasibility study is on assessing feasibility – not publishing results. Participating HEIs have however received an institution report presenting their results along international benchmark reference points.
- ² In 1989, Messick defined validity as an “integrated, evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores” (Messick, 1989).
- ³ These Standards cover six types of validity: construct validity, content validity, face validity; concurrent validity, predictive validity, and consequential validity. Given its limited timeframe, the AHELO feasibility study was not designed to provide assessments of predictive and consequential validity.
- ⁴ A construct is a representation of the underlying student ability, also referred to as latent trait. One of the assumptions underlying the instrument development for the AHELO feasibility study is that there is a latent trait which can be represented by a continuous variable and is possessed by students. Items are developed to require students to use this trait in responding to items and the amount of the trait possessed by students is a function of the score they receive on the test.
- ⁵ For the AHELO feasibility study, the Rasch Item Response Theory (IRT) model was used to analyse data.
- ⁶ See Volume 1 for more detailed information about the AHELO feasibility study instrument development process.
- ⁷ For item production, the Technical Standards (AHELO Consortium 2012) specify that any item clustering used for reporting must have a reliability estimate of at least 0.80 at the individual level. Items should have an average discrimination around 0.50, a goodness of fit above 0.95, and mean square residual (fit) statistics above 0.80 and below 1.10. Across key groups item demand estimates should be within 95% confidence limits.
- ⁸ A decision was made not to remove any of the two constructed-response tasks and their three dimensions (analytical reasoning and evaluation, writing effectiveness and problem solving) due to the significant amount of testing time they represent.

- ⁹ In the charts, item difficulty indices from female students (vertical axis) were plotted against the item difficulty indices of male students (horizontal axis). Each dot represents one item. When an item lies on diagonal line, it means that the item difficulty for females is not different to the item difficulty for males and there is no DIF for the item. An item deviating from the diagonal indicates that there is a difference between males and females in terms of item difficulty. The further away from the diagonal line, the larger the difference. Items sitting outside the 95% confidence band, illustrated by the two lines outside the diagonal line, show significant DIF. An item sitting below the diagonal indicates that the item is relatively easier for females, and vice versa.
- ¹⁰ The DIF analysis conducted across countries examines the difference between the item difficulty parameter of a country and the average difficulty of the item of all countries. An item difficulty from a country with a significantly higher index than the average difficulty of the item from all countries indicates that the item is deemed to be harder than expected in that country, and vice versa.
- ¹¹ One country with less than 100 students participating was not included in the country DIF analysis.
- ¹² The DIF analysis conducted across languages examined the difference between the item difficulty parameter for a particular language and the average difficulty of the item of all languages. An item difficulty for one language with a significantly higher index than the average difficulty of the item from all other languages indicates that the item is deemed to be harder than expected in that country, and vice versa.
- ¹³ The table summarises the number of items having at least one test language shown harder or easier than expected. Again, it is possible for an item appeared harder than expected in one or more test languages, and easier than expected in other test languages.
- ¹⁴ One country with less than 90 students participating was not included in the country DIF analysis.
- ¹⁵ The scope of the feasibility study did not call for detailed psychometric review of construct validity and reliability of these three contextual dimension surveys, which were designed for the most part to deliver discrete variables that required minimal scaling or aggregation. Although it can be said that the three contextual dimension surveys achieved content validity through being mapped onto the validated Contextual Dimension Assessment Framework, for which international consensus was obtained.
- ¹⁶ To investigate whether the items are measuring a common trait across all countries, each item was examined in terms of fit to the model (using the Rasch Item Response Theory (IRT) model). Goodness of fit to the IRT model for individual items focused on weighted mean square statistics.
- ¹⁷ The psychometric results underlying these conclusions were not available to the OECD.

- 18 Cognitive interviews were conducted with 52 students across participating countries. These cognitive interviews were conducted to collect feedback from students on their linguistic and cognitive interactions with the test materials (see Table C4 for the follow-up interview questions).
- 19 Students who participated in focus groups were also invited to provide verbal feedback as well as completing a brief questionnaire. Results from this feedback process are given in Table C5 which gives the percentage of students who agreed or strongly agreed to evaluative statements about the test material.
- 20 This was not so much of a problem for the economics instrument for which only two countries (Egypt and the Slovak Republic) joined after the expert group met. This was more of an issue in the engineering strand in which there were more latecomers (Abu Dhabi, Canada-Ontario, Colombia, Egypt, Mexico, the Russian Federation and the Slovak Republic).
- 21 Based on the scoring scheme used for the generic skills constructed-response tasks, there was no item-level missing data for the constructed-response tasks as scorers provided a score for each dimension.
- 22 Students were asked about how much effort they if they put into the test— either ‘little or no effort’ (1), ‘some effort’ (2), ‘close to my best effort’ (3) or ‘my best effort’ (4).
- 23 Students were asked to rate how relevant the test materials were to their current degree and to future professional practice—either ‘not at all’ (scored 1), ‘very little’ (2), ‘some’ (3), ‘quite a bit’ (4) or ‘very much’ (5).
- 24 Student satisfaction with educational provision is widely used as a proxy of the quality of higher education. While not exactly an academic criterion, students’ satisfaction with their educational programme towards the end of an undergraduate path can serve as a proxy indicator for academic success. The examination of how AHELO results compare with students’ ratings of their entire educational experience can provide some degree of concurrent validation of the AHELO feasibility study scores.
- 25 For the AHELO feasibility study, reliability indices are calculated using plausible values and final plausible values. Plausible values are random numbers drawn from the distribution of scores that could be reasonably assigned to each individual. Plausible values are based on student responses (plausible values), as well as on other relevant and available background information (final plausible values). For details on the uses of plausible values, see Mislevy (1991) and Mislevy et al. (1992).
- 26 Classical test theory conceptualises reliability as a function of individuals’ responses to items. When this relationship is conceptualised in a hierarchical framework, it can be

extrapolated to relations between different hierarchical levels. Reliability can then be conceptualised at the institution level by using students, as opposed to items, as measurement units clustered within institutions.

27

Four statistics were used to assess inter-scorer reliability within countries:

- The first statistic is the average difference (disagreement) between two scores (\bar{X}_d), calculated as the average of absolute difference between the two scores. The standard deviation of this metric ($s(\bar{X})$) is also presented.
- The second statistic is the percentage of absolute agreement ($\%_A$) between scorers.
- The third statistic is intraclass correlation (ρ), a measure of agreement that takes into account the degree of absolute disagreement between scorers such that a low correlation is a result of large absolute disagreements, and a high correlation means small absolute disagreements.
- The fourth statistic is perhaps the most conventional measure of scoring reliability kappa (κ).

Each of these metrics has its advantages and disadvantages. While percentage absolute agreement is intuitive it is likely to be an overestimate of 'deliberate agreement' between two scorers because unlike kappa it doesn't partial out agreement due to chance. At the same time, kappa has limitations given the structure of the data at hand. Good reliability is indicated when \bar{X}_d and $s(\bar{X})$ are both low, and when $\%_A$, ρ and κ are high.

28

For the generic skills strand, all students' constructed responses were double-scored.

29

For the economics strand, approximately 20% of students' constructed responses were double-scored.

30

For the engineering strand, approximately 20% of students' constructed responses were double-scored.

31

The mapping of item difficulty indices onto the scale illustrating the distribution of the student ability shows how well an instrument is targeted to the student population taking the test. The test is said to be well targeted when the average item parameter is approximately the same as the average estimated student abilities (Figures A13, A14 and A15).

32

There was no item-level missing data for the constructed-response tasks for the generic skills instrument as scorers were instructed to provide a score for each one of the three dimensions.

CHAPTER 8

NATIONAL EXPERIENCES

Seventeen countries/economies took part in the AHELO Feasibility Study. We have asked them to reflect on the experience. This interesting feedback is provided below, country by country. The first page for each country is the poster which was prepared for the AHELO feasibility study Conference.

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.



www.oecd.org/edu/ahelo

Key data on participation

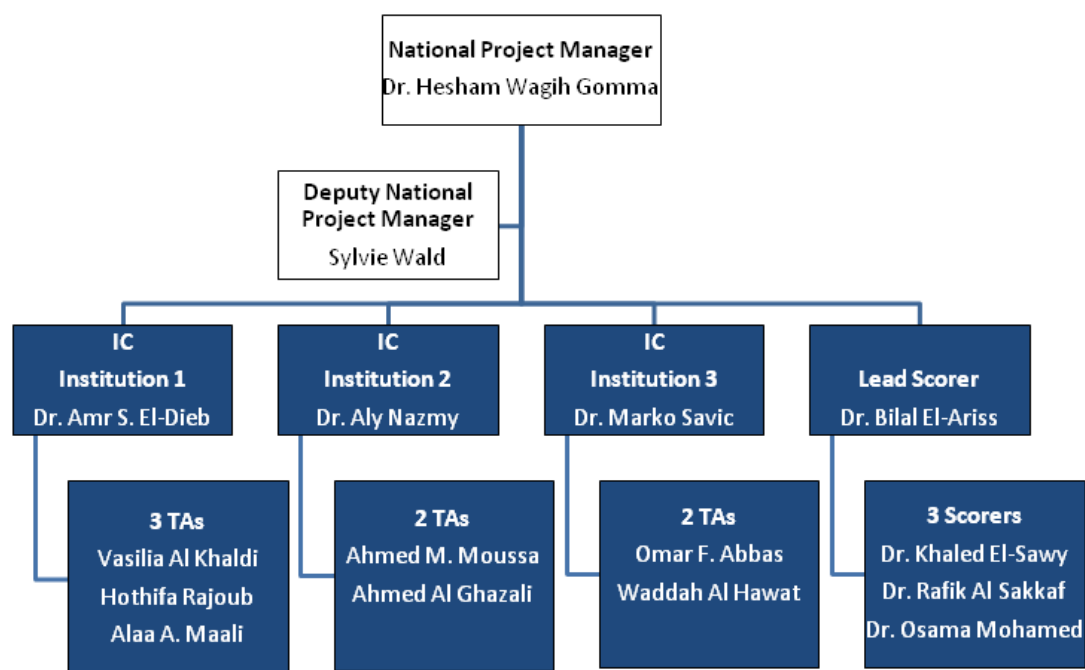
Abu Dhabi Emirate participated in the engineering strand at quite a late and critical stage of the project. As English is the language of instruction in civil engineering programs, it was decided to retain English as the testing language which allowed for valuable time to be spent more efficiently on the adaptation of instruments and implementation of activities. Out of the four invited HEIs that have civil engineering programs in Abu Dhabi, only three institutions offered the Bachelor of Civil Engineering programme and were hence eligible to participate in AHELO. In coordination with these institutions, the National Centre managed to recruit all the 135 final year enrolled students and the 44 full time faculty to participate in AHELO. As a result the response rates were 90% and 73% for the students and faculty respectively.

National and institutional management

As per the OECD/AHELO Consortium suggested project organisation structure, Abu Dhabi Education Council (ADEC) served as the AHELO National Centre (see figure 1) that included the following roles:

- National Project Manager (NPM)
- Deputy National Project Manager (Deputy-NPM)
- Three Institution Co-ordinators (ICs) one for each institutions
- One Lead Scorer (LS) and three Scorers.
- Two/three Test Administrators (TAs) for each institution.

Abu Dhabi National Centre Organisation Structure



Preparation for fieldwork

Following the enthusiastic response from the three institutions and the set up of the project organisation structure, the National Centre of Abu Dhabi developed an ambitious implementation plan to complete the project activities and meet the international timeline. The plan covered all activities such as team recruitment, kick off, training, communications, sampling, IT needs assessment, testing setup, testing implementation, and scoring, supported by quality assurance at each stage.

IC recruitment and training

Given the criticality of the project timeline, the National Centre developed rigorous terms of reference for the ICs and requested institutions to appoint faculty with senior roles in the institution such as Deans of Colleges or Heads of Departments. Meanwhile, the National Centre adapted the AHELO training material and instruction manuals, and developed an action plan to fit the Abu Dhabi context and the remaining timeline of the project. Soon after, the National Centre organized a training session for the ICs and agreed upon all proposed activities according to deadlines and key performance indicators.

TA and Scorer recruitment and training

In return, the ICs responded efficiently in recruiting the TAs and nominating highly experienced civil engineering faculty as Scorers, from whom the Lead Scorer was selected. To speed up the implementation process, the National Centre took full responsibility for the training of both Scorers and TAs, who attended training sessions and were in regular contact with the National Centre.

The sampling process

As per the Consortium's standards, "a systematic equal probability sample of 200 students [was to] be drawn from the implicitly stratified lists of in scope students" (AHELO Sampling Manual, p.12). Given the small population size (135) of all in-scope students, the National Centre agreed with OECD and the AHELO Consortium to proceed with a census instead. Opting for a census saved a significant amount of time and effort in designing the student and faculty sampling plans.

Similar to student enrolment, the number of full time faculty in Abu Dhabi is quite small. As a result the National Centre proceeded with a census for the faculty as per AHELO Consortium standards. The National Centre hosts a robust Education Management Information System (EMIS) that covers all institutional data, which assisted in defining in-scope faculty after further validation with ICs.

IT standards and challenges

One of the key success factors for Abu Dhabi is the well developed IT infrastructure and the availability of highly skilled personnel in this field. The assessment of the IT facilities in the participating HEIs showed full compliance with AHELO requirements. In addition, the NPM liaised with other participating countries and attended different real testing sessions in Egypt to identify potential challenges before testing in Abu Dhabi. The outcome of this exercise led to an efficient implementation of testing sessions.

Quality Assurance

Due to the tight deadline, the National Centre put emphasis on developing and devising a rigorous quality assurance system to enhance the outcomes while simultaneously aiming for cost efficiency in implementing the project. This made the participating institutions accountable for their performance with adequate autonomy to be dynamic and creative. The system was based on monitoring the progress made against well defined implementation norms and targets in alignment with the Consortium manuals. In case of deviations, the National Centre acted rapidly to inform institutions, and ICs reacted promptly with the necessary correction measures.

Fieldwork operations

Given the small number of enrolled students, the National Centre was keen on aiming for a recruitment rate of 100%. The National Centre based its communication strategy on the benefits of participation which emphasized international and local visibility and standing of the participating institutions, national pride, and individual incentives among other elements. The ICs

played a key role in executing this strategy. Following the training session held in the National Centre, fieldwork within the institutions required the rapid mobilization of faculty, students, and institutional research offices, which ICs accomplished under tight deadlines. Each IC arranged workshops within his institution to orient all stakeholders with the AHELO requirements. Some of these workshops included the presence of the university management which reflected the level of institutional commitment. Also, some ICs acquired board approvals to grant additional credits to participating students.

Meanwhile, the National Centre held a press conference that was attended by the institutions' Vice Chancellors, where the launch of the project was announced and covered by the national news agency, in addition to major national newspapers.

Feedback from students/faculty

Feedback from students varied drastically from the test being too long, excessively difficult, and conducted at an inopportune time to being more than manageable and an exciting opportunity to be exposed to the relevant knowledge and skills identified by the international civil engineering academic community. Faculty equally expressed an array of reaction from apprehension in regards to wide-scale standardization of the programme curriculum to appreciation for the ability to understand their students' performance and by extension, comparative teaching quality at their institutions.

Scoring process

The scoring process was the most challenging phase of the project for two main reasons. The first reason is that the AHELO testing and scoring process took place while recruited Scorers (who are also faculty members) were fully engaged in finalizing end of term commitments at their home institutions. The second reason, given the limited time for training, some of the Scorers were not comfortable about the relevance of the AHELO test to the civil engineering programs offered at their institutions. This paralleled the concerns raised during the Lead Scorers training session in Paris. This in general has enriched the discussion and scoring process outcomes.

As a result, the National Centre tried to reduce the number of days allocated to the scoring process. The scoring process was divided into 2 phases: piloting and scoring. During the pilot, each Scorer was allocated a number of items to be marked by the Scorer to be more oriented to the online scoring platform and to estimate the required time to complete the scoring for the allocated items. This process was led and overseen by the Lead Scorer who re-scored at least 20% of the items to monitor consistency as per AHELO's scoring manual.

Results

As Abu Dhabi engaged in a census, the student response rate at the institution level varied from 77% to 100%, and represented 90% of the total population. The institution reports are currently being reviewed by institutions and preliminary feedback was very positive, with institutions indicating interest in pursuing analysis at a scale larger than within their institutions. Institutions also indicated that they would be using the results, alongside feedback

already received from the multiple players involved in AHELO, to hold discussions surrounding their teaching practices.

Impact at national/institutional/faculty level

A preliminary analysis of AHELO data confirmed to policy makers that assessing only input and processes as is currently done does not give a full picture of institutional quality. At the sector level, there are discussions on the feasibility of conducting similar assessments of higher education learning outcomes that cover a wider spectrum of programmes. The National Centre is pursuing further analysis on the background of the participating students to evaluate the impact of the current education reforms in basic education.

Institutions deemed the tool useful in providing evidence about their global standing. Some institutions have started to review their programmes and curriculum to identify their strengths and weaknesses. Further follow up in the form of engagement and surveys is planned to better gauge institutional feedback to the AHELO findings.

Any particular innovative process you would like to share

Abu Dhabi had the unique opportunity, due to its late participation, to benefit tremendously from the shared experiences of other countries. This allowed the National Centre to conceive of an evidence-based risk management plan to identify the most effective implementation strategy in the local context. The National Centre developed a separate, operations-based, procedural manual in visual format, which consisted of slides grouped by activity that concisely outlined protocols required for that activity, to which was also appended a one-page checklist of key milestones, both adapted from the AHELO manuals. NPM recommendations from the March meeting were also heeded and the plan also indicated risk of inconsistencies in the reporting of institutional data and resulting inefficiencies in consolidating the information. Hence, standardized templates for quality assurance purposes were developed and explained in the training sessions. The advice was invaluable in that newly mobilized participants could proceed with implementation easily following the simplified steps outlined, while having the necessary reference documentation available if further detail was required. This also reduced the burden on ICs to filter through information in the manuals, and the templates reduced institutional workload to implementation only. In the meantime, the National Centre was able to allocate more time to focus on assuring the quality of each activity.

Any particular challenge or problem you met

Due to the late participation, the National Centre did not have the time to recruit an administration team to support the project management at the national level. Also, the National Centre team was engaged in a wide range of other major projects locally.

Suggestions for a main study

Given the diversity of sub-specialties in the higher education programmes (particularly in Engineering), it is recommended that the OECD and participating countries examine which level of sub-specialties AHELO should assess for international benchmarking in the future. It would

also be particularly informative to expand the number of discipline-specific strands to broader fields.

A specific message from your country

"If deemed feasible, AHELO may very well be the new gold standard in Higher Education and will serve as a rich student learning outcomes-focused complement to other global institutional quality measures such as ranking and accreditation."

– Dr. Mugheer Khamis Al Khaili, Director-General, Abu Dhabi Education Council

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.



www.oecd.org/edu/ahelo



Australian institution co-ordinators at AHELO Symposium, October 2011

Australian participation in the AHELO Feasibility Study: Overview of activities and outcomes

By Daniel Edwards, Australian National Project Manager

Introduction

Australia participated in both Phase 1 and Phase 2 of the Assessment of Higher Education Learning Outcomes (AHELO) feasibility study. Participation has been funded by the Australian Government. Australia chose to be involved in the Civil Engineering Strand of the study, and over the two phases of the work, eleven universities have had direct involvement in the project. Phase 1, involving focus groups and qualitative feedback on the draft instruments, attracted the participation of ten universities, gathering insight from 78 final year civil engineering students. Phase 2, involving the administration of the AHELO assessments, was participated in by eight universities, and collected data from 187 students and 87 faculty members. About 40 people across the country had direct involvement in the administration and implementation of the AHELO feasibility study.

National management

The Australian participation in the AHELO feasibility study was co-ordinated by the National Project Manager (NPM), Dr Daniel Edwards, from the Australian Council for Educational Research (ACER). The NPM was substantially assisted by Ms Eva van der Brugge, as well as other support staff at ACER. Funding for the NPM and the AHELO activities was provided by the Australian Government through the Department of Education, Employment and Workplace Relations (AHELO Phase 1) and the Department of Innovation, Industry, Research and Tertiary Education (AHELO Phase 2). The NPM and team held regular meetings with the department/s involved in the project and provided monthly progress reports during busy phases of the work.

The ongoing operation and implementation of the AHELO feasibility study also involved institutions, scorers and adaptation experts. The NPM directly liaised with institution co-ordinators, the Lead Scorer and with academics providing specific feedback on the draft assessment instruments both in terms of content (Phase 1) and adaptation (Phase 2). Institution Co-ordinators were responsible for teams within their university, in particular test administrators and student recruitment. The Lead Scorer was responsible for working with the scoring team that was gathered together for the scoring sessions by the NPM.

Fieldwork preparation

In both Phase 1 and Phase 2 of the feasibility study, preparation began with a formal invitation letter being sent to university Vice Chancellors to participate in the project. The choice of schools was determined by the Australian Government and the Australian NPM in consultation with the Australian Council of Engineering Deans. Institutions that had been active participants in previous government-level conversations about the assessment of learning outcomes and AHELO were prioritised in selection. The institutions invited were generally representative of the diversity within the Australian higher education sector.

All ten institutions invited to participate in Phase 1 took part in the focus groups and providing feedback. In Phase 2, three of the original institutions were unable to devote further resources

to the project. One additional institution was invited to join the project, and agreed, resulting in eight participating universities for Phase 2.

In preparation for Phase 2, a national meeting was held by the Australian NPM in Melbourne in late October 2011. All participating institutions were represented at this meeting, where an overview of the project was provided and suggestions for planning of the implementation of the student assessment were offered. At this meeting, institutions were asked to begin a range of tasks including: nominate an institution co-ordinator, identify appropriate testing dates and begin to inform students about the study.

In early 2012, all institution co-ordinators were sent an Australian AHELO Test Administration Manual. This manual was adapted from the AHELO Consortium's international manual to suit Australian institutions. The Australian NPM also conducted two online training sessions in early March 2012, in order to prepare institutions for the sampling of students and staff, and to ensure all were aware of AHELO test administration protocols. In addition, these sessions served as an opportunity for institutional co-ordinators from different institutions to share best practices in maximising student engagement and response rates. In between these formal sessions, continual contact was maintained between the NPM and institution co-ordinators.

The provision of sampling frames to the NPM was a significant effort on the part of the institutional co-ordinators, since it required up-to-date information on the enrolment and leave status of students and staff members, and a range of demographic characteristics – all required at a time that was the start of the academic year. The NPM co-ordinated the collection of sampling frames from the institutional co-ordinators, and processed frames to ensure adherence to the international AHELO format. The NPM then communicated with the consortium to receive login details for each institution at each participation level, that is to say, for students, staff, institutional co-ordinator and test administration assistants needing to access the AHELO online test system.

Fieldwork operation

Test administration in Australian universities took place during April and May 2012. For the administration of the student assessment and questionnaire, most institutions organised a number of sessions, while some of the smaller institutions organised a single session. In some institutions, the institutional co-ordinator supervised all test administration sessions, whereas others hired test administration assistants financially supported through the NPM budget. During test sessions, the NPM was available for support regarding the AHELO test system and any other queries regarding the AHELO protocol. The NPM attended a student test session at one institution to monitor progress directly and was in daily contact with institution co-ordinators during the test administration period.

In the vast majority of cases, test sessions ran smoothly with no issues relating to the online system. The NPM was informed by universities of the times for each session and was available to help resolve any issues that arose. Overall, most issues that arose from the online testing were simply resolved and usually occurred as a result of a step in the administration manual being missed. In one institution, where notable problems occurred during testing, the diagnosis

revealed that it was due to an issue with the configuration of the computers being used within the laboratory rather than a system/AHELO-level issue.

A major challenge for the Australian implementation of AHELO was motivating students to participate in the study. This was a problem also highlighted by some other countries during the international NPM meeting in October 2011. A range of different incentive options and approaches were offered by universities. In some institutions, the number of students was small enough to offer a voucher to each participant. In others it was necessary to have a draw of prizes for participants. One university chose not to offer any monetary or “prize” incentives for students and instead chose to focus solely on the experience of the assessment as being an incentive for participants, while at the same time integrating the assessment into a unit of study. Interestingly, it was this institution that had the most success in attracting participation.

Feedback

In general, the institution co-ordinators were positive about the processes and implementation approaches provided by the AHELO Consortium and the support offered by the NPM and staff.

While motivating students to participate was a key problem for most of the Australian institutions participating, the feedback from those students who did take part in the assessment was, on the whole, positive. For example, the NPM attended a “thank-you” luncheon at one university held for students who participated in the project. During this event, students spoke about how the assessment had stimulated their thoughts regarding how the things they were learning in their degree related to the kind of work they would be undertaking upon graduation, and further on in their careers. Students found the test challenging as well as stimulating, with many indicating that there were sections of the assessment that made them realise how much they had forgotten of some of the fundamental issues covered in earlier years of their course.

Students also indicated to institution co-ordinators and the NPM that the format of the AHELO assessment was relatively unique in their experience. First, the constructed response tasks were singled out as particularly different from their traditional thinking about what an assessment involved. Second, the online implementation of the test as a whole was a new experience in assessment for most students.

Scoring

Scoring took place in late May 2012 at ACER, following completion of testing in all universities. A team of four scorers, under the supervision of Lead Scorer, Professor Roger Hadgraft, scored all responses over a period of two days. Scorers were recruited via institutional co-ordinators and were doctoral students in civil engineering from two of the participating universities.

Australia took part in international scoring studies to further the investigation of feasibility of international comparability in scoring standards. The Lead Scorer scored an additional set of translated responses from other countries and Australia undertook a small cross-scoring project with Canada.

Outcomes, conclusion and future considerations

The Australian participation in the AHELO feasibility study has resulted in a number of valuable lessons. First, it has shown that the Australian sector is equipped to participate in this type of international study. There was interest from stakeholders in participating, and administrative systems in institutions allowed relatively straightforward production of the required student and staff data.

Second, participation in the process of developing and trialling an internationally applicable engineering test has proven to be insightful both to participating institutions and students.

Third, the co-operation between engineering experts from multiple countries both in meetings and via online communication has provided an excellent opportunity to strengthen international bonds. In addition the co-operation between participating institutions opened up opportunities for future co-operation across engineering schools.

Fourth, AHELO has shown that Australian students are not easily motivated to participate in a voluntary test or questionnaire. However, the successful engagement of students in one Australian institution, where 98% of students took part, shows that near universal participation in such activities is possible in the Australian context. Though participation rates in the feasibility study for Australia were disappointing, the process of implementation has built substantial knowledge on the processes and systems needed for engagement among students and institutions in future studies of this kind.

For Australia, there are some worthwhile considerations for future international participation that have become apparent through the feasibility study. One is that in the testing window used in this phase Australian students were technically one semester behind those from institutions in other countries (except Japan). For accurate international comparisons, future iterations of the study should be implemented in comparable times during academic years across all countries. A second is that a number of the institutions involved in the Australian participation have substantial final year internships or research projects in the final year, meaning that students spend significant time in this year off-campus. As such, being able to find a time in which a large cohort are able to participate in a secure assessment is difficult for institutions. Longer term planning for the running of such assessment could help in minimising the impact of these key events in the final year. However, the importance of the flexibility of internships and research projects may make this a challenge to achieve. A final issue recommended for consideration in the future is the production of student-level reports for individuals who participate. It is recognised that this was beyond the scope of the feasibility study, but Australia believes building such capabilities in the future would help to stimulate engagement of students in these types of studies.

AHELO has provided some small insights into the current state of engineering bachelor education at selected universities in Australia. While this data does not yet allow any strong conclusions about the skill level of Australian students as it compares to that of students in other countries, it offers a glimpse into what could be possible in future iterations of such studies. Importantly, involvement in the AHELO feasibility study has provided Australia with valuable lessons and models for implementation of such assessments in the future.

Belgium - Flanders

**BELGIUM (FLANDERS)**

- ☒ Economics
- ☐ Engineering
- ☐ Generic Skills

The AHELO feasibility study showed us that an international test to assess higher education learning outcomes can be developed.

Main Challenges

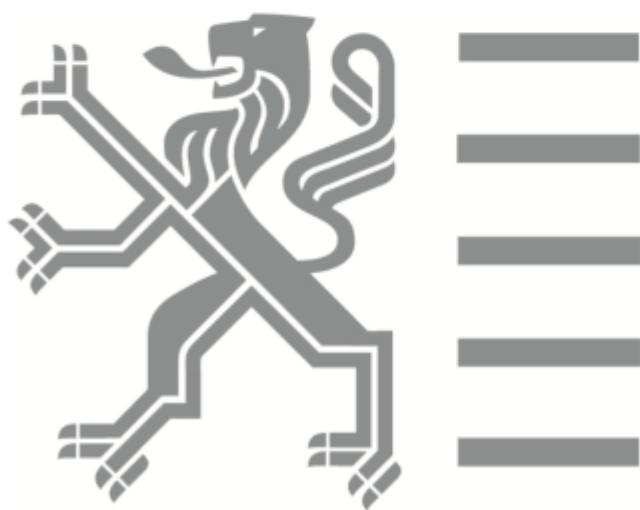
- ✓ Students were not very eager to participate. Some institutions couldn't find enough students to participate and had to stop the project. For the test we had to use a census methodology.
- ✓ The study was sometimes considered as another evaluation on top of the existing ones.
- ✓ Convincing university colleges to participate, next to research universities.

Main achievements

- ✓ Knowing that it's possible to develop an instrument for the international comparison of achieved learning outcomes.
- ✓ Learning about the do's and don'ts of developing this kind of assessment.

Main Lessons

- ✓ To find out what resistance exists among students and teachers towards these types of evaluation.
- ✓ Open ended questions should be kept to a minimum.
- ✓ It is better not to take a sample, but rather to use census methodology.



ahelo
Feasibility Study

www.oecd.org/edu/ahelo

Key data on participation

- Strand: Economics
- Number of participating HEIs: 2
- Number of students: 328
- Number of faculty: 50

National and institutional management

- National Team
 - Noel Vercruysse, Chair
 - Tony Keuleers, Member
 - Raoul Van Esbroeck, NPM
 - Eduardo Cascallar, National Expert
- Coding team
 - Iris Vanaelst, Lead coder
 - Ilse Steel, coder
- Expert team for translation and adaptation instrument

Name	Role	Qualifications	Experience
Erwin Ooghe	Involved in adapting the instrument (from the Dutch version to the Flemish version)	PhD economics	Professor economics (micro & macro) at KU Leuven
Inge Demeyer	idem	Master Educational Sciences	Expert in test construction. Was also involved in PISA
Carine Coppens	idem	Master Economics	Teaching economics at Hogeschool Gent –co-ordinator quality assessment
Jan Schelstraete	idem	Master economics	Teaching languages at Hogeschool Gent
Guido Erreygers	idem	PhD economics	Professor economics at University Antwerp – Chair department of economics

The team to support the adaptation of the instrument was recruited by the Ministry of Education through a call for candidates sent to the participating universities and university colleges. We opted for one language specialist (since it was not really translating, but rather

adapting from Dutch to Flemish). We also asked for one specialist in test construction as well as professors teaching economics. The team members were selected by the NPM, the National Expert (Dr. Cascallar) and Mr. Tony Keuleers (representing the Ministry of Education). The NPM, in association with the National Expert, supervised all the activities and participated in the team meetings. The team remained the same through the entire procedure.

- Institutional Co-ordinators
 - Universiteit Gent: Luc Van De Poele
 - Universiteit Leuven: Luc Sels and Erwin Ooghe

Preparation for fieldwork

- Training IC: face-to-face conversation at the HEI with the NPM
- Training test assistants (TA): one session per HEI (2 to 3 hours)
- IT planning: conversation per institution with the IT-manager
- Sampling: there was no sampling, since HEIs knew beforehand that not enough students would be willing to participate. Therefore, with the endorsement of ACER, we opted for a census.

Fieldwork operations

- The two HEIs took the test in a different way:
 - In Ghent, the test was part of a course in which all students had to participate and the rector called upon the students to take part.
 - In Leuven, the dean, student representatives and teachers encouraged students to participate on a voluntary basis. The response rate was low and so there are doubts as to how representative the results are.

Feedback from students/faculty

- Following phase 1, we received the following reactions:
 - “The test is too long, we didn’t get enough time to answer all questions.”
 - “The test is given at the wrong moment, i.e. at the end of the bachelor programme, which is too late. Most of the questions related to topics we’ve learned in the first or second year of the programme. If we could have rehearsed the courses, the results would have been better.”
 - “The open questions aren’t evident, you don’t know what to answer. You can answer nearly every question in different ways.”

However, results show that the students found the questions relevant to their current degree.

Scoring process

- Most difficulties were encountered when scoring the open ended questions. The lead scorer had to recode the questions (up to 3 times). This was because too large inter scorer differences arose.
- Scores for open ended questions could differ a lot when scoring either very strictly or less strictly.

Results

- The reports allow us to compare, but more specific analyses within the different options within a programme aren't possible (for example applied economics, business administration, etc.)

Impact at national/institutional/faculty level

- Since the feasibility study was only meant to investigate if an international assessment can be developed, the results of the tests can't be used for these purposes. The data obtained cannot be used to draw conclusions as they aren't reliable.

Suggestions for a main study

- When developing a main study, the framework should be prepared beforehand and the items should be linked to it. It could be used to compare an institution with similar institutions in other countries, not to compare countries as such. A fully developed assessment could help in the international benchmarking of institutions (which is needed for the accreditation process in Flanders).
- The test to be developed should be focus more on content. There should be as little testing of generic skills as possible.
- Open ended questions should be kept to a minimum and should have a lower weighting. They are more time-intensive.

Conclusion

The goal of the project is interesting and testing if the learning outcomes are achieved is an important issue. The AHELO feasibility study was only meant to see if an international assessment can be developed, no conclusions can be made from the results of the assessment as such.

Extending the AHELO study to other programmes could be desirable, but one has to realise that what was developed so far isn't a final product; we still don't have a test. The test that was used cannot be used again and needs changes and adaptations. New item collection has to be done and a new pilot has to be run.

Institutions can also use the results for curriculum reflection and reform. Compared to secondary education, university colleges and universities are more autonomous in their programmes. Differences in performance can be due to differences in curriculum planning.

Canada (Ontario)

☐ Economics
☒ Engineering
☐ Generic Skills



CANADA (ONTARIO)

Participating in AHELO significantly furthered our understanding of learning outcomes assessment, both through our own experiences and those of our international colleagues.

Main Challenges

- ✓ Student recruitment was a significant challenge requiring considerable time and effort by the institutions.
- ✓ Institutional concerns about student-level data collection meant we were unable to capture accurate information on the characteristics of the student sample.
- ✓ The timing of the assessment window was challenging as it occurred during a period when students were on study break, and writing final reports and exams.


Main achievements

- ✓ Considerable 'buy-in' from institutions, with 9 out of 10 eligible institutions volunteering to participate because they wanted to know how their institutions compare.
- ✓ Faculty members took ownership of the project and supported student recruitment because they wanted to know how their students were doing.
- ✓ Scorers found value in the assessment and said that the experience with AHELO will change their future teaching and assessment methods.

Main Lessons

- ✓ Engagement of key members within the faculty was central to successful implementation.
- ✓ Weekly Institutional Co-ordinator conference calls were valuable in keeping the project on track and created a supportive community.
- ✓ It would have been beneficial to work with institutional administration much earlier in order to ensure that all aspects of implementation were as successful as possible.

Key message: Taking part in the international assessment was particularly valuable for faculty members engaged in the implementation and scoring processes. It provided them the opportunity to reflect on their curriculum design and delivery, and on the assessment techniques they employ. Most significantly, it made faculty members question their own methods and re-evaluate what they require of students.



The Department of Civil and Environmental Engineering

[Contact](#) | [Site Search](#) | [Login](#)

OECD-AHELO Test
UNDERGRADUATE
GRADUATE STUD

Chair's Message

News

Research

About

Prospective Undergraduate Students

Prospective Graduate Students

Information for Current Students

Course Outlines

Forms

Faculty

Staff

Job Opportunities

John Adjelesan Lecture

Contact Us

Home / OECD-AHELO Test

OECD-AHELO test

Last name *

First name *

Student number (Please enter the last 3 digits only) *

Do you plan to write the OECD-AHELO test? *

☐ Yes, I am eligible to graduate next June and will write the test

☐ No, I will not write the test because I am not eligible to graduate next June

☐ No, I am eligible to graduate next June but will not write the test

Please check all the times during which you can write the test: *


☐ Monday March 12: 10:30-12:30

☐ Monday March 12: 5:30-7:30

☐ Tuesday March 13: 12:00-2:00

☐ Tuesday March 13: 5:30-7:30

☐ Wednesday March 14: 10:00-12:00


www.oecd.org/edu/ahelo

The province of Ontario, Canada, joined the Civil Engineering strand of the AHELO feasibility study in July 2011. The Canadian National Project Office for AHELO is housed within the Higher Education Quality Council of Ontario (HEQCO), a research agency of the Government of Ontario which was asked to conduct the study on behalf of the Province. Ontario is the only Canadian province that joined the feasibility study, although other provinces are keenly interested in it and have been kept apprised of the process through the Council of Ministers of Education, Canada. A number of national agencies have also been following the AHELO activities, including Human Resources and Skills Development Canada, the Association of Universities and Colleges of Canada, and the Canadian Engineering Accreditation Board. Sharing the activities and engaging in discussion with a broad range of stakeholders was seen as an important element of success in the eyes of the National Office.

Ten institutions in Ontario provide bachelor's degrees in Civil Engineering. HEQCO invited each to participate in the feasibility study and offered them a small amount of funding to cover the basic costs of administration and implementation. The response to HEQCO's call was overwhelming, with nine of the ten institutions immediately agreeing to participate in the project. The institutions noted their interest in taking part in this international assessment as a way of understanding their own programme, those next door, and those a world away through a comparative lens. Given that the demographic makeup of students and faculty in Canadian Civil Engineering programs tends to be particularly international, participation in AHELO offered institutions an opportunity to support mobility by better understanding the characteristics and knowledge base that exist in other countries.

The participating institutions are representative of Ontario's universities. All are public institutions that offer a broad range of arts and science programs up to the doctoral level. Located primarily in urban areas, they range in size from 14 595 to 75 941 full-time equivalent (FTE) students. Participating institutions include:

- Carleton University
- McMaster University
- University of Ottawa
- Queen's University
- Ryerson University
- University of Toronto
- University of Waterloo
- Western University
- University of Windsor

The Civil Engineering programmes are housed in Faculties of Engineering, and are occasionally partnered with Environmental Engineering. The programs have between 17 and 40 faculty members, 90% of whom are full-time staff and 98% of whom hold doctorates (according to

AHELO survey results). The FTE student populations in the programmes range from 231 to 573, and the number of undergraduate degrees awarded ranged from 20 to 101 in the 2010 academic school year. Demographically, 94% of students in the participating programmes are under the age of 25, 79% are male and over 90% pursue their studies full time. Approximately 90% of all Ontario's Civil Engineering students are represented in these programmes.

Following initial interest in the AHELO project from the participating institutions, there was a flurry of activity to ensure that we would be able to implement the test in time. The Lead Scorer took responsibility for vetting the tests with faculty and students to verify compatibility with Ontario's curriculum and technical language. Minor changes were made as a result of this process, but there was agreement that the assessment was suitable for Ontario students both in terms of content and difficulty. The National Centre reviewed the context surveys in collaboration with the Lead Scorer and modified them to ensure that students, faculty and institutions would be able to provide the most accurate information possible. This review provided a good reminder that each nation is unique in its system structure and institutional organization.

Recruiting Institutional Co-ordinators (ICs) was a straightforward task. In most cases, the Chair of the Civil Engineering departments volunteered for the role. Their leadership was vital to the success of the project, and became a key aspect in securing the engagement of departments, faculty members and students within their institutions. The ICs also took on the significant role of working with the institutional research ethics boards. Although the executive administration and the individual departments agreed to participate, each Ontario institution independently determines what research can take place within its walls (including research conducted by or involving its faculty and students) based on national research ethics codes.

The research ethics boards had reservations about the use of student-level data, citing concerns about the fact that they could be linked administratively to the individual student (even if only by the IC), and that the data would be housed outside of Canada. Based on our tight timelines, we altered the research proposal slightly in order to gain institutional ethics approval. The modification meant that we were not able to create population frames or identify students in any way and are now unable to determine how representative the sample is of the Civil Engineering student population.

While more of an administrative than an ethical issue, it is very rare in Ontario for institutions to mandate that their students write an administrative test, especially during class time. As a result, the IC's had to be creative in approaching faculty and students to engage in this low-stakes test.

Faculty were informed about the study and recruited either during staff meetings or through informal discussions in some of the smaller institutions. Faculty members' appreciation of this study can be seen in the high faculty survey response rate (72%). Having the support of faculty members at each institution was important for information dissemination and successful recruitment of students as well. In some cases, the professors allowed short presentations explaining the AHELO project during class time, and they themselves promoted the assessment

to their students in other cases. Such reminders and recommendations to students were no doubt helpful for recruitment.

Student recruitment presented perhaps the largest burden on the ICs and their teams. Encouraging participation from students who are in the last term of their programme, completing final reports and projects, and studying for university and professional exams can be extremely difficult. The remarkable response rates of 48 to 78% can be attributed to the creative student recruitment strategies developed at each institution. Many institutions used a combination of common recruitment techniques such as posters, classroom presentation, or emails. One institution created a website to inform students about the study and encourage them to sign up. Similarly, a range of incentives were offered to participants, ranging from small monetary amounts to raffles for larger monetary or other prizes. Some institutions chose to work through student peer groups. In a few cases, the Engineering student societies were provided a nominal sum of money for their support with recruitment, and the amount would increase if they managed to recruit a certain percentage of their peers. At another institution, the monetary amount provided to each participating student would increase if the stated recruitment rate was met. Hence the administrative burden of recruitment, which was found to be quite significant by some ICs, was reduced in a few institutions. Overall, the institutions recruited approximately 61% of their students to write the test. This accounts for nearly 60% of all the graduating Civil Engineering students in Ontario.

Despite the high recruitment numbers, the actual test numbers are lower. As is understandable in feasibility studies, there was a technical challenge at one institution. Based on an unforeseen issue with the internet server, one university lost 92% of the student data. This was frustrating to the institution, which would no longer receive any data on its students despite having had a very successful administration. It, like the other institutions, had tested the IT capacities and was satisfied that it had conducted appropriate troubleshooting to prevent any potential issues. When the problem was noticed during test administration, representatives contacted the National Project Manager, who contacted the Consortium. The error, which resulted in the loss of 7.4% of Ontario's results and dropped our response rate from 61 to 58%, was a disappointment both for the institution and for our broader jurisdictional work.

Despite the disappointment of knowing that they would not be receiving their results, this institution's commitment to the AHELO project was intact. When the call for Scorers was presented, the IC volunteered to take part in the process. He, along with five other civil engineers from across the institutions, participated in a two-day scoring session held in Toronto where 2 463 responses from Ontario students were graded. Based on arrangements made between the Canadian and Australian National Centres, the Scorers also graded a small number of Australian student responses. As this was their first opportunity to see the assessment, their first comments noted that generic skills were not addressed in the strand assessment.

The opportunity to discuss and examine the questions asked and the answers provided proved to be the most interesting part of the process for the ICs. They were often surprised by the responses that students gave, noting that they as instructors do not ask students to think in the same way that the test did. They liked that the test was trying to ascertain the 'above content',

the application of knowledge. The scoring process led the faculty members to reconsider both what they were teaching their students and also how they were assessing them.

This reflective process points to one of the most interesting aspects of the AHELO feasibility study in the Ontario context. An international test that can provide another way for programmes and faculty members to think about teaching, learning and assessment methods is of significant value. In this case, only a small group of Scorers had the opportunity to learn from the AHELO results. In a wide-scale assessment, and one where the data is provided in a way that can broadly inform programmes and institutions, the impact can have a further reach.

As the results come in and publication begins, we expect to provide a range of analyses. The institutions have their reports, but are looking forward to seeing the jurisdictional analysis in order to properly contextualise the information. We plan to develop a comparative analysis with other jurisdictions based on mutual agreement and data sharing. The institutions have noted that they would specifically like to see a report exploring all aspects of the Civil Engineering strand in a comparative lens. Given that this goes beyond the scope of this feasibility stage, we recommended that future work in AHELO ensure there is suitable data collection to allow for analysis at the strand level.

With respect to the main study, the Ontario experience suggests that there would be value in incorporating some generic skills tests into the strand-specific assessment. This is a good opportunity for the programmes to learn as much as possible about their own students, and also provides a benchmark that can be used comparatively across programmes. A further suggestion would be to open the test window so that it accommodates various system-level structures, in order to capture students at the end of their programme, and at a time that is administratively feasible.

Colombia



COLOMBIA

☐ Economics
☒ Engineering
☒ Generic Skills

Taking part in an AHELO feasibility study has been of great significance for all of Colombia's different Higher Education stakeholders and in particular for ICFES, the Colombian Institute for Educational Evaluation. Important challenges were met and valuable lessons were learned along the way. All in all it was a great opportunity to take part in a discussion at the highest level about the technical and practical requirements of a Higher Education assessment.

Main Challenges	Main achievements	Main Lessons
<ul style="list-style-type: none"> ✓ Short timelines both to organize the application and to complete the marking of constructed response tasks. ✓ Finding an application scheme to ensure the possibility of studying the relationship between SABER PRO and AHELO. Given Colombia's unique position in which all of its higher education graduates take an end-of-degree test, participating in AHELO provided a unique opportunity to compare the results of both tests and enrich the AHELO data with information collected for the national test. 	<ul style="list-style-type: none"> ✓ Selection of participating institutions and programmes was successful: 25 of the 26 programmes approached decided to take part in the assessment. ✓ Nearly 4 000 students assessed in one day in 26 application sites across 18 cities. ✓ High students' response rates: due mainly to the strategy to couple the application of AHELO with that of SABER PRO: the median for Generic Skills was 95%, with minimum 91%; and for Engineering the median was 98% with minimum 79%. 	<ul style="list-style-type: none"> ✓ Allow more time for Faculty responses. ✓ Devote more work to discuss and adapt test items and marking grids.

Key message: Linking the results from AHELO, SABER PRO (including the socio economic data) and SABER 11 (end of high school exam) will help evaluate the possibility of producing higher education value added measures. Besides pursuing its own analysis, ICFES intends to make these data public and to provide support to interested researchers.





www.oecd.org/edu/ahelo

Members of the National Team

- Julián P. Mariño (GNE Member)
- María Camila Perfetti (NPM)
- María José Figueroa (Generic Skills coding leader)
- Andrés Guzmán (Engineering coding leader)
- Julián Segura (Application co-ordinator)

Introduction

Taking part in an AHELO feasibility study has been of great significance for all of Colombia's different Higher Education stakeholders and in particular for ICFES, the Colombian Institute for Educational Evaluation. Important challenges were met and valuable lessons were learned along the way. All in all it was **a great opportunity to take part in a discussion at the highest level about the technical and practical requirements of a Higher Education assessment.**

Colombia's approach to the AHELO feasibility study was considerably different from that of the other participating countries. For the whole study, the institutional and management organization of the test application was highly centralized. ICFES used its existing national structure to centralize the co-ordination of the 36 participating HEIs. In this way, many of the sampling, co-ordination, test administration and training tasks were executed directly by the National Centre.

Colombia participated in both the Generic Skills and the Engineering strands. The tests were administered as part of the national end-of-tertiary-education exam, called SABER PRO. In total 36 HEIs, in 18 different cities, were selected to participate in AHELO. The number of sampled students reached 4 034 for both strands (3 000 in GS, 1 034 in ENG). In addition, 1 253 faculty members were sampled for the Faculty Context Instrument.

Challenges

Colombia encountered two **crucial challenges** for the successful implementation of AHELO:

- **Time shortage both to organize the application and to complete the marking of constructed response tasks.** The decision to participate in phase 2 of the Generic Skills strand of the study was made only in March 2012. Colombia was thus among the last countries to deliver the test and, as a result, the deadlines to mark the performance tasks and the constructed responses of over 4 000 evaluatees were very tight. This demanded a great effort from both scoring teams. The Engineering team, made up of 7 Scorers, managed to complete the task on time, whereas a time extension had to be requested to allow the Generic Skills team, with 13 Scorers, to complete the scoring of over 6 000 performance tasks.
- **Finding an application scheme to ensure the possibility of studying the relationship between SABER PRO and AHELO.** Given Colombia's unique position in which all of its

higher education graduates take an end-of-degree test, participating in AHELO provided a unique opportunity to compare the results of both tests and enrich the AHELO data with information collected for the national test.

Achievements

Ultimately the AHELO feasibility study came through and the great efforts of ICFES, HEIs, faculty and students involved paid off. The following **key achievements** were attained:

- **Selection of participating institutions and programmes.** For the Generic Skills strand, from among the institutions with more than 200 registered students for the SABER PRO application, a representative group of 15 were chosen. Then, from each of those institutions, 200 students were randomly selected to participate in the application (excluding those enrolled in civil engineering programs). For the Engineering strand, all the Civil Engineering programmes with more than 20 registered students for SABER PRO were invited to participate with all their registered students. Only one programme, out of 26, declined to take part.
- **Nearly 4 000 students assessed in one day in 26 application sites across 18 cities.** The AHELO application was programmed to take place on 2 June, the day before the SABER PRO national test application. Computer rooms in 26 different SABER PRO application sites (all Higher Education Institutions but not all of them participating in AHELO) were previously inspected, had their technical features verified and were successfully set up for the application. Despite fears of breakdown because of the high concurrency, no major problems were encountered and the test administration was successfully completed.
- **High students' response rates.** The strategy to couple the application of AHELO with that of SABER PRO was the main reason behind the very high student response rates in Colombia: the median for Generic Skills was 95%, with minimum 91%; and for Engineering the median was 98% with minimum 79%.

Major issues

During the entire process **major issues** became evident, some of which are worth mentioning to improve future implementation processes:

- **Need to devote more work to discuss and adapt test items and marking grids.** It was observed that some items were not appropriate for the Colombian context. Once they started working with real students' answers, the teams that had to score the constructed responses and the performance tasks encountered difficulties that had been overlooked in the adaptation process.
- **More time required to get faculty answers.** Response rates from Colombian faculty were low. More time to diffuse and then monitor and react to this problem would have helped reduce its impact.

The very tight schedule to complete the study did not allow for a proper solution of these problems.

Message

The administration of AHELO jointly with SABER PRO and the big size of the samples involved open important roads for research, of which two are of special interest. As mentioned above, a first study of particular relevance for the country's assessment system will be to explore the relationship between both exams. This will help identify lessons to improve the SABER PRO tests. Second, by linking AHELO results with those of the high school exit exam (SABER 11, which is compulsory to enter higher education) and with the socio economic data collected in the SABER PRO registration form, the set will be complete to evaluate the possibility of producing, out of AHELO results, higher education value added measures. Besides pursuing its own analysis, ICFES intends to make these data public and to provide support to interested researchers.

Suggestions for a main study

Three important lessons have been learned in Colombia's 10 year long experience with assessing Higher Education learning outcomes. These seem to be valid for any large-scale assessment in the higher education sector:

The first lesson relates to what can be assessed. Here, the deep specialization of higher education poses a major challenge. The original approach of ICFES' exams was to develop specific assessments for many different careers. On the one hand, this was very costly and, on the other, it resulted in a multitude of measurements with little coherence. This proved to be of little use at levels of the higher education governance structure above the programme or the department.

To address these difficulties while preserving a wide scope in the assessment, the only alternative is to focus on skills and competencies that are shared by large groups of careers. There is widespread agreement that higher order thinking skills such as Quantitative Reasoning, Critical Thinking or Written Communication are examples of "universal" aims of tertiary education programmes for which there are successful cases of assessment development. Examples of competencies that are not "universal" but still relevant to many different careers are Scientific Thinking (for scientists as well as engineers and some health professions), Pedagogy and Assessment (for education students), Health Promotion and Disease Prevention (for health students) or Project Management (for different business oriented students). ICFES has decided to work on the development of assessments of these kinds of learning outcomes. Another alternative would be to have assessments for different areas such as Natural or Social Sciences, Health, Business, Education, Agriculture, etc.

Any such approach will have to deal with criticism of its inability to measure the most distinctive competencies that each particular higher education program is designed to breed into its students. However, it is clear that any evaluation through a standardized large scale assessment (and particularly an international one) is enormously restrictive in terms of formats

and content. Therefore, it cannot attempt to be “the quality measurement” but must restrict its aim at producing relevant indicators to inform the decision making processes.

The second lesson has to do with how to set up the different tests. If one intends to evaluate both generic and more specific skills, it is intuitively appealing to have specific tests that incorporate the assessment of generic skills within the specific subject. However, this strategy has a major drawback. It precludes the possibility of producing comparable results in cases where such comparison would make sense and be truly useful. If a common skill is assessed using different items with different populations the results are hardly comparable. To avoid that problem and be able to produce comparable results, common skills must be identified and tested with common items. This cannot be done within specific subjects. That means that generic skills assessments should be kept independent. They should not be incorporated into the more specific assessments.

Finally, there is an important insight into the level at which results of Generic Skills should be reported and analyzed. The fact that generic skills are “universal” does not mean that one should expect all tertiary-education students to reach the same level of mastery. For instance, one should expect, and this is fortunately the case, engineering students to do on average better at Quantitative Reasoning than law students, while the latter do better at Written Communication. This is also true for semi-specific competencies such as Scientific Thinking, where physics students tend to do better than engineers. As a consequence, the overall averages on generic skills of Higher Education institutions depend on the composition of students from different programmes which is rarely similar from one institution to another.

The relevant comparisons between institutions should be made using averages of groups of similar programmes rather than overall averages. It is fair to compare the overall average scores of the engineering programmes of any two institutions, whereas it would not be fair, and can make little sense, to compare the overall averages of a business oriented institution with those of another specialized in health or education programmes. Hence, a classification of programmes has to be established, grouping similar programmes, among which one can fairly compare generic learning outcomes. Results should be produced for each such group within the different institutions. Comparing results across such groups can be meaningful and very useful but should always be done bearing in mind the difference in programme orientation.

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.



ahelo
Feasibility Study

www.oecd.org/edu/ahelo

GS AHELO team in Translation Adaptation Training

Prof. Dr. Ibrahim Shehatta

NPM, AHELO EGYPT

Egypt has contributed successfully to the success of the AHELO feasibility study in spite of all the obstacles and instability the country is currently facing. The zeal, spirit and good will created by the revolution, as well as the enthusiasm and motivation of all individuals working in and with the project team (NPM, National team, scorers, ICs, TAs, institutions' teams, students and faculty), provided a substantial driving force to face the challenges, with the aim of improving Egypt's competitiveness in the global knowledge-based economy. Also, the continuous support of the higher education authorities (ministers of higher education and rectors of universities) and the valuable assistance of the AHELO Consortium and the OECD, which was greatly valued at this juncture, meant that the AHELO implementation in this leading MENA country was noticeably successful.

Key data and information

Egypt participated in all three strands of the AHELO feasibility study: the Generic skills, Economics and Engineering strands. A total of **19 universities** have implemented the three tests, representing governmental and private universities that are geographically distributed all over Egypt, with a total participation of 4 212 **Students** and 877 **faculty** from various academic programmes (both accredited and non-accredited).

The AHELO study was entirely funded by the government through the Ministry of Higher Education.

The central management and local centres

An AHELO-EGYPT National Centre (NC) was established within the context of the Egyptian National Centre for Measurement and Assessment in Higher Education. All activities were conducted through networking and continuous communication between the NC in Cairo and local centres at each of the 19 participating universities (via emails, letters, phone calls, face to face meetings, seminars and workshops).

The National Project Manager (NPM) was responsible for the implementation of the AHELO feasibility study at national level – ensuring that all required tasks were carried out on schedule and in accordance with the prescribed technical standards and operational guidelines – and for documenting processes implemented at the national level. The NPM followed a teamwork management style and applied principles for open door management and de-centralisation of action planning and implementation. He recruited qualified members for the team and assigned tasks according to individual experiences and capabilities.

The NC conducted a risk assessment and drew up contingency plans to deal with the many possible risks/threats, such as: national status, commitment of students, fulfilment of IT requirements, unexpected incidents during test implementation, etc.

Preparation for fieldwork

Awareness campaign

Three cycles for AHELO awareness were conducted (December 2010-January 2011, October-December 2011 and February-April 2012). This included brochures, media announcement, booklets and guidelines, seminars and focus meetings. More than 250 000 students and 15 000 faculty staff were made aware of AHELO.

The awareness campaign was organised on several levels, over several events and targeted at different categories of stakeholders, for example:

- General conference with the Minister of Higher Education, the presidents and vice Presidents of universities, the deans of faculties, staff members (faculty) and student representatives.
- Seminar by the National AHELO Team at each of the 22 governmental universities in Egypt
- General meeting with ICs of the different strands, representatives of staff members from different programmes participating in the AHELO survey, as well as student representatives.
- Separate focus meetings for each strand **separately**: a meeting was held between the national team of each strand and the ICs of universities participating in a particular strand, together with representative samples from staff members and students who expressed their deliberate intention to participate in the contextual survey and in the tests. The meetings aimed at giving a general orientation regarding competencies to be measured, as well as a general work plan for the particular strand.
- Technical strand specific meetings with ICs, representative samples of selected students and staff members.

In addition, the national team frequently visited the selected 19 Egyptian universities and held discussions about AHELO with selected students and staff members.

Training

Training was an asset all through this study: several face-to-face and online workshops were implemented. The national team was keen to acquire new skills and competencies through training and exchanging experiences with the AHELO Consortium, as well as the other participating countries, aiming at transferring such new concepts in assessment and evaluation to the participating universities' faculty and students.

Repeated rounds for training workshops targeting ICs, Test Administrators (TAs), IT teams and institutions' teams were conducted for the successful implementation of AHELO online tests.

Training of selected Students (12-24 April 2012):

- For the Economics and Engineering Strands tests, students got acquainted with the test outline, the content, the test instructions, and the general procedures for scoring.
- For the Generic skills strand, students recognised the concept of the performance task, and trained using the translated mini PT “reduction of traffic accidents caused by using mobile phones during driving”.

Sampling

As regards sampling, two major challenges appeared and threatened the compliance with AHELO guidelines: the first was the large number of programmes and students in the Egyptian universities; the second was the incomplete electronic databases of students in most universities. This entailed formulation of a specific strategy – “Egypt’s Sampling Strategy” – which was extracted from the AHELO sampling manual, and reviewed and approved by Statistics Canada in the AHELO consortium.

The sampling strategy followed a stratified, non-random pattern and described the selection criteria for each target:

- For universities: selection was done according to the number of programmes and fulfilment of the requirements for the IT infrastructure.
- For students: selection was based on gender, high school education and grade of achievement during the pre-final academic year (excellent, very good, good, passing).
- For faculty: selection was done according to gender and academic ranking.

IT planning

IT planning required reviewing all computer labs in Egyptian universities. The selection of a university as a study participant was linked with its fulfilment of AHELO IT specifications and facilities. Around 150 test administrators (1 per 40 students) and 150 professional IT technicians (2 per lab) were involved through the use of nearly 100 computer laboratories.

An effective IT strategy was developed and implemented to carry out the AHELO online tests concurrently in 19 Egyptian Universities. This strategy followed the prerequisites stated in the AHELO IT Manual and consisted of the following fundamental domains: computer labs, internet services, security systems and management technique.

Fieldwork operation

The high-level support from national authorities was a great help for the implementation phase of the AHELO tests. The Minister of Higher Education invited the rectors of selected universities to:

- promote AHELO/Egypt as a national priority towards reform and improvement of the higher education system;
- co-operate with the AHELO/Egypt national team;

- provide all the facilities and finance needed to fulfil the requirements for the AHELO tests;
- approve the budget allocated for incentives for the universities' teams (ICs, TAs and IT personnel).

The date and time of test sessions were set to avoid conflicts of students' course schedules and periods of heavy use of the universities' internet networks. For each strand, the ten participating universities implemented the test sessions concurrently on the same day over ten cities all over Egypt.

The NPM received from the AHELO Consortium the actual and spare usernames and passwords for students, faculty and ICs, and distributed them to the responsible Institutional Co-ordinator (IC) at each university who, in turn, managed communications with the selected students via emails and mobile messages.

On the days assigned for the three strand tests, ten of the well-trained National Team Members travelled to the ten universities all over Egypt to monitor the test procedures, to reply to inquiries from ICs, TAs and the IT team and to solve any emerging problems by direct communication with the central main control team in Cairo.

Feedback from students/staff members (faculty)

The following quotations represent samples from students and staff members, who participated in AHELO tests and surveys.

Students

"It is a great experience for me to share in an international study like AHELO."

- Students participating in the generic skills test:

"The Performance Task is a non-traditional test, we need more training on similar tests."

"The Performance Task is a positive exercise, engaging, concise, not tedious and almost fun."

"The Performance Task is an incredible assessment; it is engaging and challenging."

- Student participating in the engineering test:

"The MCQ items are very interesting, I wish our teaching and assessment could be changed to that direction".

- Student participating in economics test:

"The ideas of the MCQ items are complicated; we are not using them during our study".

Staff members (faculty)

"We will be looking forward for the analysis of the results. Hopefully it will highlight the defects in our current education."

"This is a fantastic assessment tool that can help us to improve our teaching/learning and assessment processes."

"I participated to be part of an objective and scientific research study that might have an impact on the process of reform of the Higher Education system in Egypt."

"These economics and engineering tests asked students to apply their knowledge and skills rather than just memorising contents."

"The performance task is truly tested cumulative analytical skills and the ability to think in a challenging and logical way."

The scoring process

The **scoring** process lasted for nearly 1.5 months because of the high response rate in the three strands. Scoring for the engineering strand test (CRT) ended on 25 June 2012, on 28 June 2012 for the economics strand test (CRT) and 10 July 2012 for the generic skills strand test (CRT).

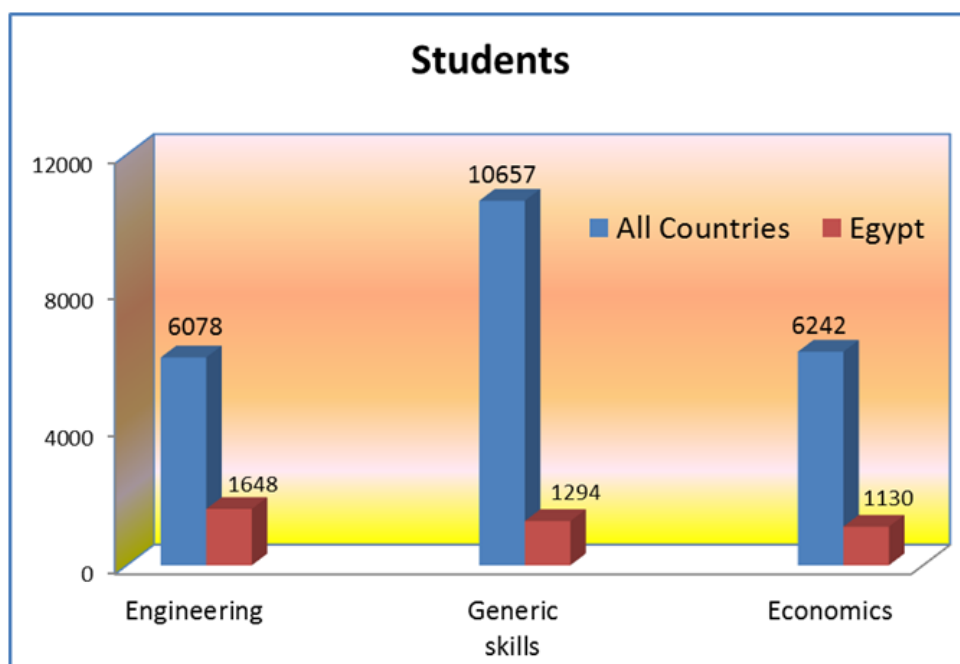
Online scoring was an interesting new experience for all scorers in Egypt. Fourteen scorers participated in scoring AHELO tests after extensive general and specialised training.

Results*Participating institutions and response rates achieved***Participating institutions and response rates achieved**

Respondent	Strand					
	Generic Skills		Economics		Engineering	
<i>Students</i> (International targeted students participation = 1500 / strand)	1 434	(95.5%)	1 130	(75%)	1 648	(110%)
<i>Staff member (faculty)</i> (International targeted faculty participation = 350 / strand)	319	(91.1%)	231	(66.0%)	327	(93.4%)

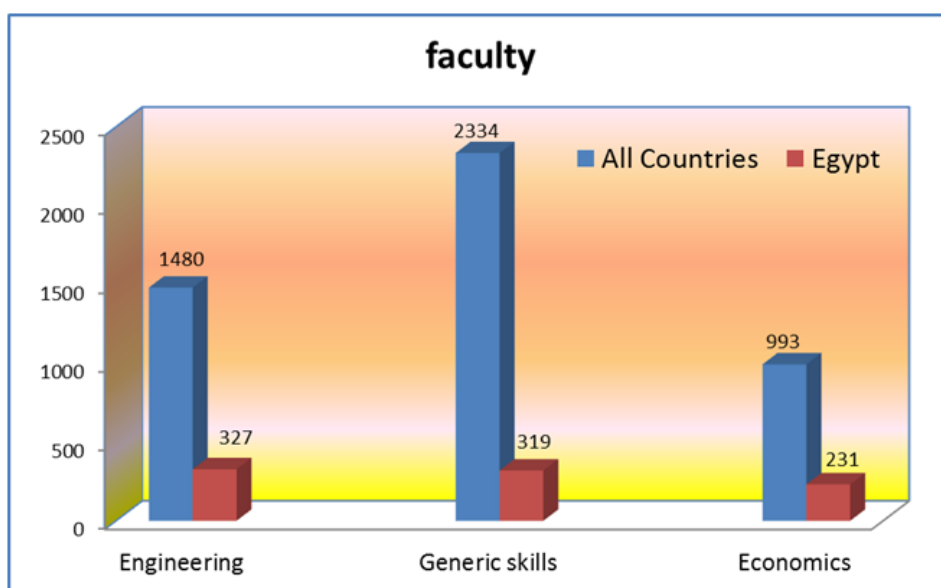
Students

The input by Egypt to the AHELO constitutes 18.3% of the total participating students.



Faculty

The input by Egypt to the AHELO constitutes 18.2 % of the total participating staff members (faculty).



Value of institution reports received from ACER

Institution reports were expected to answer important diagnostic questions and so illustrate the weaknesses/strengths in the educational environment, thus helping to formulate action plans for correction, improvement and enhancement. However, the AHELO institution reports that were received provided general results without deep analysis. It did not meet the high expectations of the institutions, which had spent considerable effort and time in managing and implementing the AHELO test. The correlation of AHELO scores (and not the skills assessed) with the contextual variables is not of great help. Hence, the participating institutions cannot make any use of the report in its current form.

Although AHELO has been designed to check the validity and applicability of the assessment items and not designed to produce internationally representative scores, the design of the assessment and the large number of participating students permits and deserves further analysis of the results to investigate which skills are satisfactory and which need reinforcement. The AHELO EGYPT National Centre will try to make further analysis to provide useful data and information to institutions. The skills mastered/missed by students need to be identified and benchmarked to equivalent populations.

Impact at national/institutional/faculty level

Egypt considers participation in the AHELO project an investment from which the profits would be expressed in the form of valuable data, its analysis and its benchmarking with other countries.

One can expect requests from the various stakeholders (government, the Ministry of Higher Education, the Supreme Council of Universities, policymakers, institutions, faculty, students, accreditation agencies, sponsors and professional organisations/syndicates) to use the obtained evidence-based data (AHELO results) for the following purposes:

- reviewing graduates' skills;
- modification of curricula and teaching methods to promote self-learning and development of students' generic skills;
- adjustment of the intended learning outcomes to fit the requirements of the labour markets, both nationally and internationally;
- focus on the assessment of learning outcomes in addition to the inputs and processes;
- evaluation of the systems for quality assurance in higher education, aiming at closing the chain of inputs, processes and outputs.

Innovative process and best practices

- Egypt's engagement despite a revolution: the NC made contingency arrangements to overcome unexpected issues in relation to the national political status and unstable conditions emerging during the fieldwork

- The central management and local centres.
- Establishment of a help desk team for supporting universities.
- IT strategy and UTM policy: repeated visits to computer labs have been made to verify fulfilment of AHELO-IT requirements, using check lists (the last check-up was run the day just before the test sessions); internet connections inside universities were suspended, except in computer labs at time of the tests to assure full and efficient service; the test time was 2 pm (end of working hours) to avoid internet overload in general.
- Applying a unique identifier coding system to overcome the deficient digital student databases in most universities.
- Hands-on training for students, by simulation, of the Constructed Response Test (CRT) for the generic skills strand, using a translated version of the released mini performance task after verification by the Council for Aid to Education (CAE).
- Monitoring of the test process: ten members of the national team were present in the ten universities while the tests were being implemented to monitor test implementation, provide support and ensure continuous contact with the national centre and international technical support for co-ordination.
- Governmental financial support for funding: participation of Egypt in AHELO feasibility study, incentives to institution teams, missing institutional IT resources and other executive activities.

Challenges and problems

- Rapid and radical changes that involved the whole Egyptian community.
- Recruitment of students and faculty to participate in the AHELO feasibility study.
- Incomplete digital databases at most high education institutions.
- Repeated changes in the higher institutional leadership and management boards that altered the schedule of implementation of the project activities.
- Achieving the HE authorities' and stakeholders' satisfaction with the outcomes of the AHELO study, specifically as regards the comprehensive data sub-dimensions analyses.
- Absence of individual competencies' scores makes further analyses difficult and limits the benefits of the AHELO results.
- How the country data and institutional reports resulted from the AHELO feasibility study would be used to explore the weaknesses/strengths in the educational environment.

- Conformity of the outcomes of the AHELO feasibility study to the original objectives and expectations that would promote further national investment in similar education research studies in the future.


Positive signs for improvement

- The universities, faculty and students showed interest and enthusiasm in participating in AHELO in all strands and activities.
- The universities showed great interest in:
 - establishing local AHELO centres at their campuses; and
 - expanding the scope of AHELO to involve more academic programs.
- Institutions and staff asked for detailed reports that illustrate the strengths and weaknesses and disseminate best practices for high-score institutions and diverse learning experiences and skills.
- Ability of Egyptian Higher Education to implement a wide scale electronic online test system.

Suggestions for a future main study

- The time frame should consider enough space for pilot studies and the training of targeted populations and of the implementation teams.
- The study design should include simulations of tests using released test instruments for training purposes, as well as for the purpose of exploration of pitfalls and how they can possibly be avoided.
- The participating countries or bodies should have greater involvement in designing and managing and decision-making for the study.
- Deeper analysis of the results to obtain clear, objective, valuable and useful data for all levels (national, institution, faculty and student).
- Developing an efficient international data centre having a load balance, high availability design and security plans that permit multiple users (up to the millions) to login efficiently at the same time and protect against test system collapse.
- Necessity for a clear description of the inter-partners rights and responsibilities in contracts for future research studies.

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.

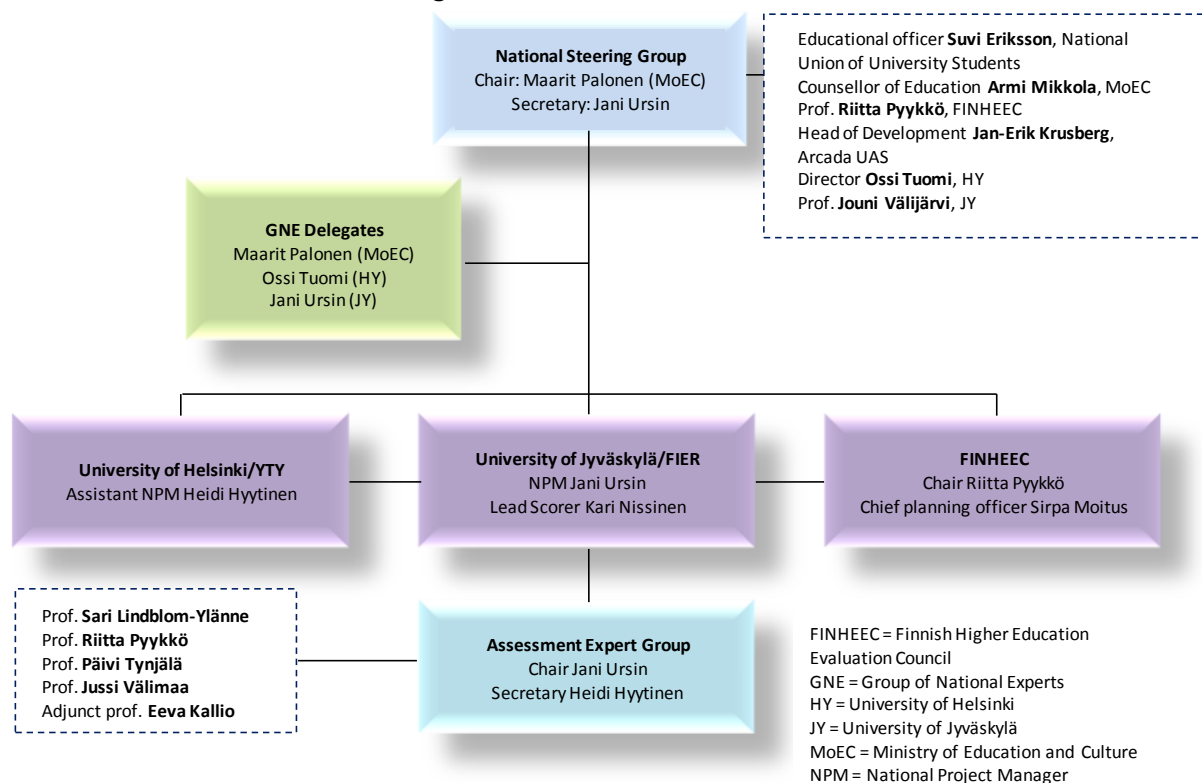


Key data on participation

Finland participated in the Generic Skills strand. Altogether 12 Higher Education Institutions, 278 faculty members and 330 students were involved in AHELO.

National and institutional management

National organisation of AHELO in Finland



A national steering group was appointed in 2009 to supervise the project in Finland. The steering group had members from the Ministry of Education and Culture, HEIs, the Finnish Higher Education Evaluation Council (FINHEEC) and the student unions. The national centre was the Finnish Institute for Educational Research (FIER) of the University of Jyväskylä, in co-operation with the Helsinki University Centre for Research and Development of Higher Education, as well as with FINHEEC. Furthermore, an assessment expert group was established to support the project.

Each participating institution also nominated an Institutional Co-ordinator (IC) with whom the National Project Manager (NPM) communicated. For implementing the field tests, each institution had a number of Test Administrators (TAs), usually either their teachers or administrative staff.

Preparation for fieldwork

Random sampling was conducted in each participating institution in co-operation with the ICs. Sampling was a smooth process, as each HEI had good and up-to-date student and faculty databases from which sampling frames were easy to establish. Before the fieldwork, all ICs were trained in face-to-face meetings in which all the preparations required from HEIs were explained and discussed. The ICs were responsible for recruiting and training TAs in their institutions. Either the NPM or the assistant NPM participated in most of the first test sessions at each HEI, thus observing and helping if needed. ICs or TAs could phone the national centre at any point during the testing. Overall, there were only minor problems – often related to IT – during fieldwork operations.

Fieldwork operations

Typically, HEIs in Finland first sent a letter – signed by the president or vice-president of the institution – to the sampled students to invite them to participate in the test. In order to give the students more choice, all HEIs organised several test sessions between February and May. In some HEIs, the sessions were distributed across a ten-week period. In all HEIs, a student could choose, via the HEI's intranet, a session that best suited his/her own schedule. Students were typically followed up by email. In one HEI the vice-president called the students and asked them to participate in the test. All HEIs also had various incentives to motivate students. Despite all these efforts the student participation rates remained poor in Finland.

Feedback from students/faculty

Some faculty members felt that not all the questions were relevant to them and some had difficulties in accessing and finishing the online questionnaire. Altogether, however, the faculty questionnaire functioned well in Finland. Although students felt that the assessment instruments were interesting, they found the test session to be too time consuming and quite challenging. Some students also complained about the lack of individual feedback from the test results. Also, for some students the internet-based test platform proved to be confusing.

Scoring process

In the Generic Skills Strand the international scorer training was adequate as regards learning the principles for assessing responses of various levels (examining benchmark papers, etc.). By contrast, the training on using the CLA online scoring system was insufficient and in some sense misleading, since it gave an erroneous picture of the Lead Scorer's possibilities to monitor individual scorers' progress and perform inter-rater reliability analysis.

In Finland we made the mistake of recruiting only two scorers per task, partly because we did not realise that the Scoring Manager was not tailored for the case of just two scorers. It appeared that the Re-score function of the system did not work and the problem was not fixed in time, which left no sensible way for the Lead Scorer to correct inadequate scorings. For a scorer, the online system worked satisfactorily and was reasonably user-friendly. For the Lead Scorer, it was too limited. The Scoring Manager manual was not detailed enough.

The domestic scorer training was probably insufficient, since in the real situation inconsistent scores were more frequent than was expected on the grounds of the experienced training. The decision to let some scorers work remotely proved bad, since it became difficult to control the scoring process, in particular due to the unexpected shortcomings of the Scoring Manager.

Results

Participation rates in Finland were low. The overall response rate for students was 13.8% (ranging from 3.5 % to 31.5% per HEI) and for faculty 58%. There were several reasons for the low response rate among students. First, the timing of testing fell at the end of spring semester when a majority of students had already left campuses to do their thesis, for internship or for employment. Second, external incentives (such as iPad and cell phone lotteries, free movie or lunch tickets) were not powerful enough to attract students to attend a test. Third, the lack of intrinsic incentives (such as study points, individual feedback from the test) further turned students off participation.

Because of the low student participation rates, the institution reports offer little information to HEIs to improve their teaching and learning activities.

Impact at national/institutional/faculty level

Some institutions have indicated that they will inform the top management of the results of AHELO and will present the main findings at internal development days. At the national level, further descriptive analyses of AHELO data will be done in spring 2013. In mid-May a national AHELO seminar will be organised in which the main results of the feasibility study will be presented and the future of AHELO discussed. All the relevant stakeholders will be invited to the seminar.

A particular challenge or problem

The biggest challenge in Finland was engaging students to participate in the test session. The main reasons for this have already been explained earlier. Additionally, students in Finland – like in other Nordic countries – enjoy great autonomy and they cannot be demanded to participate in tests like AHELO. Therefore, in a fully-fledged AHELO new (intrinsic) means – like study points – to motivate students need to be considered. This would also call for further input from HEIs.

Suggestions for a main study

From the perspective of Finland, the following suggestions can be made:

1. International financing of the project has to be fully secured before it can start.
2. The international consortium has to have a solid and consistent understanding of what kind of instruments to develop and how to carry out such a large-scale international comparative project like AHELO.
3. Enough time must be reserved for the implementation phase; more time is needed to motivate students, to train ICs and to organise test sessions.

Italy



ITALY

☒ Economics
☐ Engineering
☐ Generic Skills

The implementation of the AHELO feasibility study in Italy was a positive and successful experience which has shown that universities want to be assessed and perceive exercises like AHELO as an opportunity and not as a threat.

Main Challenges

- ✓ Highly time consuming activities for all the individuals involved in the project at international and local level.
- ✓ Continuous necessity to adapt procedures during the implementation of the project due to the experimental nature of a feasibility study.
- ✓ A general lack of familiarity in the Italian culture to have students take standardised tests.

Main achievements

- ✓ Introduction of the first experience of learning outcomes assessment in the Italian higher education system.
- ✓ Successful implementation and local management of the entire exercise, proven by the high number of universities that applied and students involved.
- ✓ The implementation of a follow up at national level on the basis of the methodology tested with AHELO.

Main Lessons

- ✓ The IT assistance during the test administration should be managed at the local level in order to allow real time troubleshooting.
- ✓ This experience has shown that Italian universities want to be assessed and perceive exercises like AHELO as an opportunity and not as a threat.
- ✓ The translation of academic content from English to Italian is a delicate and challenging task.



www.oecd.org/edu/ahelo



©Luis Rosa/Tickr

The interest raised among Italian universities by the economics strand of AHELO was remarkable. Even though only ten institutions were selected for the final test administration, there were initially 32 universities, out of the 52 which deliver courses in economics in Italy, that expressed their desire to take part in the exercise. The high rate of institutional participation resulted in the involvement of more than 1 000 students in the country. Such figures were achieved through the successful management of the process, in addition to:

- The extensive activity carried out by the National Project Manager (NPM) and her national office, aimed at promoting the importance of the project among academics and other relevant stakeholders.
- The opportunity, presented to universities, to use the test results as an incentive to students to participate, by awarding them extra credit depending on the score achieved compared with the average score reported at institutional level.

The National Centre team was made up of the NPM, Fiorella Kostoris, the assistant NPM, Massimo Carfagna, and Bruno Losito, an expert in assessment methodologies who supported the National Centre in the first year of activity.

A preliminary phase of activity consisted of a series of meetings between the National Centre and representatives of the Italian Ministry of Education, as well as other national experts, in order to discuss the methodological aspects necessary to correctly implement the AHELO procedures.

An extensive survey was carried out to draw a precise picture of the Italian university system as regards the Higher Education Institutions who deliver courses in economics. Data were collected in terms of geographical distribution, number of students, legal status (public/private) and, with an in depth analysis, the specific educational content of the single courses in economics.

At the same time, the National Centre gathered all the information available about previous and ongoing experiences of Italian universities in submitting questionnaires to students in order to identify the best procedures to fit the AHELO test.

An official message was sent to all the university Rectors to invite them to participate in AHELO by appointing an Institutional Co-ordinator. Thirty-two out of 52 universities agreed.

Several meetings were arranged between the National Centre and the Institutional Co-ordinators (ICs) to describe the initiative, to discuss the methodology and to plan the activities.

A webinar with ACER allowed the National Centre to provide ICs with specific training aimed at conducting focus groups. The focus groups, carried out in June 2011, represented an important moment that anticipated certain relevant aspects of the process: in particular, the organisational efforts and the high level of commitment required of the ICs and the institutions, on the one hand, and the high degree of difficulty of the test for the students, on the other hand.

The translations for the questionnaires were another big challenge faced by the National Centre, given the specific academic content of the test. The Italian team received useful support in this task from the valued contribution of cApStAn.

After the selection of the ten institutions that would administer the test, the ICs began the important tasks of student recruitment, organising information campaigns, identifying suitable venues for the test equipped with computers with online access, and training Test Administrators and IT staff. The overall process of test administration in Italy was fully successful, despite the National Centre not being provided with some of the information regarding the technical procedures. However, only a few problems with online connection to the system occurred at some universities.

The scoring procedures went smoothly as well, thanks to the ten scorers appointed by the ICs (in some cases the ICs also covered the role of the scorer). Unfortunately, it should be underlined that, in this case also, the instructions received from ACER for the system management were sometimes missing or misleading.

Another problem was that the possibility initially allowed by ACER to make use of the students' scores as an incentive turned out to be unfeasible and few and insufficient data were made available to the National Centre and, in turn, to Institutional Co-ordinators. Unfortunately, ACER was not able to overcome this obstacle, leading to the inevitable loss of credibility of the National Centre with the Institutional Co-ordinators and, in turn, of the latter to their students.

In conclusion, however, the implementation of the AHELO feasibility study was a positive and successful experience overall.

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.



www.oecd.org/edu/ahelo



The Japanese Scoring Team (June 2012)

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.

A total of 504 students and 196 faculty members in twelve Higher Education Institutions participated in the engineering strand of AHELO. In May 2012, when the implementation took place, the students had just started their fourth and final year in their civil engineering programs. Faculty members consisted of the entire team of full time professors and lecturers who were responsible for the education of the targeted students. The institutions were public (8) and private (4), of varying size, and from around the nation, all with capacities to confer bachelor degrees in civil engineering.

The Japanese AHELO team

In February 2009, the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) commissioned Research on the OECD-AHELO feasibility study (Japanese AHELO Team, or Team, hereafter). Chaired by Professor Kikuo Kishimoto of the Tokyo Institute of Technology, and consisting of professors and administrative staff from twelve higher education institutions, as well as experts in engineering and in higher education research, the Team conducted research on university quality assurance systems and student mobility schemes around the world. This activity, consisting of international field research, lectures by guest speakers, and hours of discussion, proved to be critically important for the implementation of AHELO in Japan for two reasons:

- Firstly, through this activity, the team came to share the sense of urgency to embrace learning outcomes based quality assurance schemes, and the idea that AHELO would provide important implications in pursuing this course.
- Secondly, the activity helped sustain the engagement of the higher education community during the prolonged planning period between 2009 and 2011. In effect, almost all of the institutions the Team members represented decided to participate in the small scale validation in May 2011, as well as the field implementation in May 2012.

Within the Team was created a small task force, which functioned as the advisory group for the National Institute for Educational Research (NIER), the National Centre for AHELO. NIER consulted this expert group for verification of instrument translation and the scoring of student responses for the small scale validation in 2011. Their inputs were crucial in improving the quality of the Japanese translation and for providing important comments to the AHELO Consortium upon modifying the scoring rubrics.

This research driven collaboration among engineering and higher education research experts proved invaluable to developing a deep understanding of the AHELO endeavour, and to generating strong engagement by individuals and institutions that care sincerely for the education of their students. The implementation of AHELO in Japan could not have happened without the commitment of this research Team, and the financial and moral support provided by the Ministry.

Struggle against tight timelines

The Japanese AHELO Team met in late March 2012, when NIER conducted an information session and invited institutions to participate in field implementation. Since our academic year begins in April, NIER spent the first weeks of April busy communicating with Institutional Co-ordinators and Test Administrators in twelve institutions to prepare for the fieldwork. We decided to conduct census sampling, partly because most of the institutions could not secure finalised versions of student rosters until mid-April, but mainly because most of our civil engineering programmes turned out to be relatively small, consisting at most of about 100 students per cohort.

Testing sessions were scheduled between 23 April and 25 May. In some of the institutions, securing computer laboratories and the necessary numbers of computers proved to be a challenge, since this was contingent upon institution-wide or department-wide scheduling. The lesson learnt here was that scheduling needed to have taken place weeks in advance, to allow for co-ordination within each institution.

The Institutional Co-ordinators were requested to designate AHELO-IDs to each student, and to submit in advance the list of student information to NIER that would then be forwarded to the AHELO Consortium for verification of the sampling frame. Some of the institutions refused to give out personal information, such as student names or Student University IDs. In such cases, a solution was reached to submit to the AHELO Consortium a list with ad-hoc student IDs and retain at NIER a list that links ad-hoc IDs to student names. An important consideration for future implementation is to formulate a policy regarding the protection of student confidentiality and to state that policy formally upon requesting for institutional participation.

A brief systems check was conducted based on the manual provided by the AHELO Consortium in each institution approximately two weeks before their testing dates. In hindsight, a full-fledged rehearsal, including opening up the testing session through the Test Administrator website and logging on to the students' testing sites, should have been conducted in order to avoid some of the systems trouble encountered during actual testing. A users' manual with information on anticipated systems trouble should be prepared for future implementation.

In terms of incentives, NIER prepared book vouchers worth JPY 5 000 (or USD 56, USD 1 = JPY 90) for each participating student.

Fieldwork operations

Several systems problems were encountered during fieldwork operations. There were instances where testing was delayed because of browser restrictions on pop-ups, which could have been avoided with better preparation. There were many instances where the computer froze, and in a few cases, information students had typed in was lost. These problems will need to be fully investigated before a main study.

Students commented that because there were so many instances of computer failure, they felt uncomfortable going back and forth on their computers to re-examine their answers. This could have been one of the reasons why many finished before the time limit and why their answers were relatively short.

Student feedback was positive. While they felt comfortable with the multiple choice questions, they also found the constructive response tasks interesting and thought provoking. They said that they liked focusing on real life problems, investigating causes of failures, proposing solutions and thinking about their responsibilities as engineers. They also pointed out that their curriculum needed updating to include more project based activities and group work, so that they will be better prepared for their careers, as well as for AHELO-like assessments.

Generating consensus through scorer training

The scoring process was overseen by Professor Kishimoto. As Lead Scorer, he participated in the International Scorer training sessions, where he contributed to the modification of the scoring rubrics. Coupled with contributions from Lead Scorers representing other participating countries, as well as the excellent management by ACER, we believe that this training process was instrumental in making the scoring rubric more focused and, at the same time, more comprehensive.

The international training was also important in the sense that experts were able to reach agreement on the logic of scoring. Because scoring requires consensus on what kinds of responses can be identified as correct, the scoring exercise urged scorers to define precisely the scope and level of learning outcomes that students are expected to demonstrate. In effect, the training generated a clearer understanding of the scoring rubrics, as well as a sense of trust that scoring would be conducted in a consistent manner across countries.

The scorer training within Japan took place immediately before scoring sessions in June 2012. Scorers consisted of twelve professors from seven institutions. Two had prior experience in scoring as task force members for the Japanese AHELO Team, and others had acted as Test Administrators. Civil engineering professors from the Tokyo Institute of Technology also responded to our request for support.

Under the leadership of the Lead Scorer, Scorers spent an afternoon and evening for training. After trial and error, it was decided that the best way to proceed was to work in two groups of six scorers. First, they would work individually on the training materials, and then work as a group to discuss why particular responses should be given particular scores. The Lead Scorer initiated the discussion and gradually shifted into a supervising role. The process was slow and controversial in the beginning, but eventually speeded up as scorers reached agreement on the general logic of scoring.

The scoring took two full days, but proceeded smoothly. For the items that were double scored, the average reliability score for exact agreement reached 89.11%, indicating that for almost 90 percent of student responses, two scorers gave exactly the same score for a response.

The innovative instruments developed by the AHELO international team were eye openers for our Scorers. They prompted the professors to reflect critically on how they teach and test their students, and inspired them with alternative approaches. The professors also discussed the importance of conducting careful analysis of the correlation between multiple choice questions and constructive response tasks, in order to deepen our understanding of what students need

to know and be able to do in order to “think” like an engineer, as well as verify the validity of our conceptual framework of engineering competencies.

Results yet to be shared and analysed

The average response rate was 65.4%, varying from 13% to 100%. The reason given for the low (13.0%) response rate was that although the event was announced to all target students, they were not strongly encouraged to sign up due to the difficulty in securing computer labs. On the other hand, response rates for half of the institutions exceeded 80%. Hence, a much higher response rate can be expected if more time is given for planning.

NIER conducted an initial review of institution reports and is in contact with ACER for clarification on several issues. A feedback meeting is planned in March 2013, where representatives of participating institutions will gather and discuss the content of the reports and future use of AHELO data.

Translating constructive response tasks

As one of the thirteen countries that implemented the assessment in a non-English language, we would like to highlight the challenges involved in translating constructive response tasks.

Because constructive response tasks are designed to “measure” how students can “think” like an engineer, their quality relies on a thoughtful balance between open-endedness and preciseness. As such, the translation of constructive response tasks was a particularly difficult task. In order to control for item difficulty, protocols of translation prohibited adding or dropping information and restricted changing the original item format, such as the order of information being presented, how sentences are divided, and how grammatical expressions were used (tense, voice, clause, etc.). This resulted at times in awkward or roundabout translation, making the items appear less straightforward and more difficult. Subtle differences in the use of technical engineering terms added to this problem. The lesson learnt here was to aim for substantive equivalence in translation, which requires some flexibility in the application of protocols, as well as the teamwork of qualified translators with extensive knowledge of the language and the subject matter.

Suggestions for a Main Study

Through AHELO, we accomplished a tangible and substantive understanding of a conceptual framework of engineering competencies and learning outcomes that can be shared globally, as well as concrete and innovative ideas for measuring competencies and learning outcomes. Suggestions for a main study include:

- Re-examine protocols for translation. Aim for a protocol that best facilitates substantive equivalence. Based on the protocol, develop a translation manual to be shared by countries using the same language. The manual should include concrete examples of difficulties encountered and lessons learnt from the AHELO feasibility study.
- Secure sufficient time and resources for the modification of scoring rubrics. The Lead Scorer training was instrumental in generating consensus on the scope and level of

expected learning outcomes. The importance of this consensus should not be overlooked, as it was the basis of consistency in scoring across countries

- Develop a system in which AHELO can inform curriculum improvement. AHELO can become a powerful tool for educational innovation, by making instruments and scoring rubrics fully available to participating institutions, and by coupling it with workshops that induce discussion about curriculum design and encourage innovation in teaching and learning.

Conclusion

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.

Korea



KOREA

☐ Economics
☐ Engineering
☒ Generic Skills

The AHELO feasibility study represents a journey towards excellence in higher education.

Main Challenges	Main achievements	Main Lessons
<ul style="list-style-type: none"> ✓ Securing sufficient budget for AHELO at the national and institutional level. ✓ Recruiting randomly sampled students across over 50 different departments at each HEIs!! ✓ Ensuring quality in scoring student answers; the scoring work was very important for the success of test implementation 	<ul style="list-style-type: none"> ✓ Randomly sampled representative data that include students and faculty from over 50 different academic fields with little governmental support. ✓ Established cooperative network among KMOE, experts of higher education, and members of Higher Education Institutions. ✓ Drew interest and concerns of stakeholders on the ways in which that the current Korean university students are educated including the issues of goals, curriculum, and pedagogy. 	<ul style="list-style-type: none"> ✓ We have witnessed mounting demand for more relevant data on higher education learning outcomes, particularly in relation to Generic Skills. ✓ We should consult with experts throughout the whole process of implementation and open communication opportunities with involved project members many as possible. ✓ Rather than other forms of incentives, authority is one of the best motivators for participants to make a commitment to the AHELO project

Key message: "Details, details, details": In order to bring about a successful "Further Studies", and serve the interests of stakeholders in the future, thorough and more elaborative schemes should be developed which include: objectives, implementation strategies, expected outcomes as well as its relevance to higher education.



www.oecd.org/edu/ahelo

Key data on participation

Korea participated in Generic Skills strand with 1 340 students nearing the end of their undergraduate degree and 170 faculties from 9 HEIs. Among participating HEIs, 5 institutions were located in metropolitan/capital area, whereas 4 institutions were in non-metropolitan/rural areas. The status of the establishment of institutions was also evenly divided into 4 national/public institutions and 5 private institutions. In terms of number of participating students, we modified the sampling size of students in order to secure the number of participants that was required for data analysis (See point on fieldwork operations for more details). With the exception of student sample size, sampling and recruiting students and faculty was undertaken according to the procedures and guidelines given in the AHELO sampling manual and IC manual. The tables below show the exact figures of participants in the AHELO test.

Number of participating HEIs, students and faculty by region

Location					
Metropolitan			Non- Metropolitan/Rural		
Number of Institutions (%)	Number of Students (%)	Number of Faculty (%)	Number of Institutions (%)	Number of Students (%)	Number of Faculty (%)
5 (55.6%)	733 (54.7%)	99 (58.2%)	4 44.4%	607 45.3%	71 41.8%

Number of participating HEIs, students and faculty by establishment status

Status					
National/Public			Private		
Number of Institutions (%)	Number of Students (%)	Number of Faculty (%)	Number of Institutions (%)	Number of Students (%)	Number of Faculty (%)
(44.4%)	706 (52.7%)	84 (49.4%)	5 (55.6%)	634 (47.3%)	86 (50.6%)

National and institutional management

The AHELO research team within KEDI (Korea Educational Development Institute) played the role of the AHELO National Centre in Korea. The Ministry of Education in Korea commissioned KEDI for the implementation and management of AHELO. Part of the national budget for AHELO was supplied with the KEDI research fund. For five years (2009-2013), a total of 15 researchers were involved in the research project of AHELO in Korea. The level of involvement of each individual varied. Three to five full-time researchers within KEDI carried out daily tasks for AHELO.

Each participating HEI in Korea appointed two to three Institutional Co-ordinators (ICs). Korean ICs can be categorized into two groups in terms of the office they belong to: the office of planning and evaluation and the office of academic affairs - usually the centre for teaching and learning. ICs were varied in terms of age, seniority, position at their institutions and work methods. Usually ICs had an internal support for the work involved in sampling, data collecting, contacting students, and co-ordinating the test administration while some ICs had to handle these tasks by themselves.

Preparation for fieldwork

To organize and manage fieldwork effectively and efficiently, we tried to work cooperatively with 18 Institutional Co-ordinators within a centralized system. We organized about six face-to-face meetings with ICs to share all directions in advance so that they became more familiar with the procedures. We also conducted several training sessions to allow ICs to closely follow the procedures required for training, preparation, test administration, monitoring, and reporting. Miscellaneous issues were also discussed through emails and phone-calls. The results of an IC feedback survey conducted in July 2012 also confirmed that the assistance from National Centre was responsive and the way of communication with the NPM was very supportive. Also, for the NPM, a centralized system was useful to ensure consistency among the different institutional contexts.

We sampled students six weeks before the dates of testing. We followed the sampling design in the AHELO Sampling Manual except for the sampling size: we increased the sampling size for each institution from 200 to 300 in order to secure a high enough number of students. Two factors were taken into account for setting the sample size at 300. First, ICs suggested that information on 150 students for each HEI should be the bottom line for meaningful data analysis at the institutional level. Second, we expected the response rate would be around 50% to 60%. Considering this, eight institutions sampled 300 students and one small university sampled 180 students instead of 300. A total of 2 580 students were sampled in Korea.

Fieldwork operations

Recruiting student and faculty was one of the main responsibilities of ICs. Various strategies and activities were carried out in order to encourage participation. At the institution level, all participating HEIs contacted student and faculty individually via email and SMS emphasising the benefits of participation in AHELO. Some institutions used their own university website for posting AHELO advertisement. Furthermore, providing monetary incentives and offering awards to students with good records (top 10%) was considered, and providing participation certificate was carried out at the level of National Centre. However, some IT system-related errors were consistently reported from all participating institutions. It should be improved before an AHELO main project begins.

Feedback from students/faculty

Regarding the instrument faculty members provided positive comments on the Performance Tasks, especially, on the ways in which the PTs assess students' generic skills. One expert pointed out that by providing a range of complex materials, the instrument aims to assess

students' skills at a more in-depth level than conventional test tools. And some faculty mentioned that this type of assessment tools should be introduced more in the Korean higher education setting.

Students found the PTs interesting, but the test length too long, especially taking two different types of tests (multiple choice questions and the PT), as well as the contextual survey.

Scoring process

One Lead Scorer and 13 Scorers were involved in scoring for 10 days including scoring training sessions. During the training sessions, Scorers had scoring practices several times until they could share an adequate level of objectivity and consistency in scoring. Scoring sessions were divided into three sessions (morning, afternoon and evening) per day. Flexibility on participating in scoring sessions was given to each scorer so that they could meet the requirements but according to their own schedule and circumstances. Scoring was conducted in one work-place, and Scorers were not allowed to fulfill scoring in other places, such as their work or home due to confidentiality and quality management.

Results

The response rate achieved in Korea was modest. The overall response rate for all Korean HEIs was 51.9%. Out of 2 580 sampled students 1 340 students participated in the test. Response rates for individual institutions varied, ranging from 37.7% to 62.3%. There was a tendency for the response rate for a public HEI to be higher than that of private counterparts. For the faculty survey, the overall response rate for all Korean HEIs was 47.2%. Out of 360 sampled faculties 170 ones completed the survey. There was a substantial discrepancy in institutional response rates; the lowest was 7.5% and the highest was 90%. In order to enhance the participation of student and faculty, there needs to be a more proactive strategy and approach in promoting the AHELO project both at national and institution levels.

The value of the institution reports received seemed to be less informative than HEIs expected. More information and comparable data on institutional analyses would also be required in order for participating HEIs to benchmark themselves against other institutions and identify areas where they can improve their performance.

Impact at national/institutional/faculty level

It seems to be too early to mention the impact of the AHELO feasibility study on national/institutional/faculty level. Certainly, most participants in this project have recognized considerable potential in AHELO as a driver for reforming higher education. For the time being, however, we have not witnessed signals that AHELO has affected changes in teaching/curriculum.

Any particular innovative process you would like to share

The National Centre (KEDI) sent one or two AHELO national team member(s) to every test setting in order to assist Test Administration, thus the quality could be controlled. Before assigning members they were asked to understand the mechanism of the test system and guidelines in the manuals. In doing so, issues and problems occurring during the session could

be resolved immediately without any confusion and trouble. ICs also expressed satisfaction with support from the National Centre.

Recruiting Scorers who were currently teaching university students provided an opportunity to share ideas on the ways in which current university education/curriculum is relevant to AHELO. An in-depth discussion between Scorers, the NPM and Lead Scorer on this issue confirmed the necessity of development of international tools for the direct assessment of student and it also provided a practical point of view on AHELO assessment tools. Moreover, scoring students responses in one work-place made it possible to have a higher quality and to maintain confidentiality.

Expanding sample size was effective in increasing the number of participants. Especially in the Generic Skills strand that targets across-discipline students it is significant to achieve the maximum number of participants in order to get meaningful analysis. It should be suggested that a modification of sample size be carefully considered in an AHELO main study.

Any particular challenge or problem you met

Securing the budget for AHELO was one of the most challenging issues. A couple of factors including the ambiguous goals of AHELO made it difficult to provide rationales for national or institutional funding for this project. In particular, officials from the Ministry of Finance and the Ministry of Education asked for unambiguous information on how to use results of AHELO and in what ways such results contribute to addressing national agendas for higher education. However, the current frameworks of AHELO seem to be limited to provide clear-cut answers to these inquiries.

To be sure, recruiting students was challenging; to reflect the voluminous and diverse system of Korean higher education, the randomly sampled students were scattered across about 50 different departments! This made it extremely difficult for ICs to co-operate with staffs at each department in mobilizing students.

Three suggestions for a main study

(1) The assessment of Generic Skills should be continued using both the PT and the multiple choice test. The CLA-type instrument of Generic Skills gained positive feedback from Korean experts for the following reasons:

- First, compared to the multiple choice test, the CLA-type instrument was more appropriate to measure high-level cognitive abilities. Korean experts also pointed out that the characteristics of CLA are more relevant to the value-added approach of assessing university students' learning outcomes.
- Second, the instrument was directly indicative of what and how to improve teaching and learning practices. Some expert argued that using only the multiple choice test would make AHELO look like a ranking driven approach.

(2) Priority should be given to HEIs when designing the assessment. Unlike PISA the AHELO project was targeted to addressing the need of HEIs. Moreover, the AHELO project was not feasible without the full commitment of HEIs. The needs of HEIs should be reflected in the plan

for the use of data from AHELO. In particular, we suggest more international benchmarks should be offered to HEIs.

(3) We should take more systemic and centralized approaches when designing sampling and recruiting students.

We have learned that authority is one of the best motivations for students and faculties to make a commitment. Monetary incentives played a limited role in creating interest to sampled students. Information offering incentives such as an individual report would be working effectively although they would incur administrative burden and obscure the goal of AHELO. Collaborative efforts between OECD, MOEs of participating countries, HEIs, and AHELO national centres can grant authority to the AHELO project.

Another important lesson we have gained is the complexity in defining target population. We have learned that diverse factors shape the definition of target populations, each of which is closely linked to the interest of individual HEIs and then this affected test results. International comparability of learning outcomes would be marred by allowing discrepancies in target populations of participating HEIs.

Kuwait



KUWAIT

☐ Economics
☐ Engineering
☒ Generic Skills

For the State of Kuwait, the AHELO feasibility study represented an international team effort at enhancing national and institutional accountability and responsibility toward student learning and success.

Main Challenges	Main achievements	Main Lessons
<ul style="list-style-type: none"> ✓ Translation and cultural adaptation of the Performance Task was difficult because the instrument chosen was based on a US tool which was not quite right to assess non-American modes of thinking or writing. ✓ The Kuwait National Team understood the difficulty associated with soliciting student participation on a national scale at the level of taking the assessment and with regards to the level of effort applied by the student being high enough that the instrument can be used as a measurement of degrees of learning accomplished by the student. ✓ Performance Task on-line platform: students were confused as to when the assessment had ended (when finishing a section for example). Thus, the assessment platform requires further development. 	<p>Initiation of a more focused and expanded national and institutional conversation on the importance and benefits associated with:</p> <ul style="list-style-type: none"> ✓ standardising assessment measures ✓ internationalising expectations with respect to student learning; ✓ internationalising benchmarks against which the quality of institutions, quality instruction, and quality learning can be measured. 	<ul style="list-style-type: none"> ✓ Need for a more comprehensive, unified and a more academically informed plan with respect to student incentive and participation, as well as to find ways to encourage "active" student engagement during testing (mandatory assessment, assessment for grades/credit). ✓ Need to internationalize the initial development of the Performance Tasks that more accurately reflect the linguistic nuances, cultural sensibilities and sensitivities, as well as student learning outcomes measures. ✓ National culture of assessment: need to mainstream standardised assessment on all institutional levels and encourage a "national culture of assessment."





www.oecd.org/edu/ahelo

Student taking AHELO test at Kuwait University

Key data on participation

The State of Kuwait participated in the Generic Skills strand exclusively both as a pilot for other future strands that the State may wish to partake in, and in order to determine the feasibility of standardising a cross-border measure that could determine the degree to which graduating students in both public and private sector post-secondary institutions satisfy the learning outcomes that students acquire through their respective general education curriculum/programmes or through their degrees, such as analytical and critical thinking, quantitative and qualitative skills, inductive and deductive reasoning, as well as problem solving ability.

The table below provides summary data on participating institutions.

Summary data on participating institutions

Educational Institute	Type of Educational Institute	Student Population	Student Sample	Number of Student Responses	Percentage of student response relative to student sample	Faculty Sample
Kuwait University	Public	34786	243	96	40%	60
PAAET	Public	14066	250	91	36%	56
American University of Kuwait	Private	2288	250	86	34%	60
Arab Open University of Kuwait	Private	6611	301	81	27%	42
Gulf University of Science and Technology	Private	3150	200	50	25%	50
Australian College of Kuwait	Private	651	185	33	18%	15

National and Institutional management

The Management of the AHELO project of the State of Kuwait required the formation of one national standing committee for the planning and organising of AHELO; one National Technical Committee (*ad hoc*), which was responsible for the translation and cultural adaptation of the performance tasks; and local institutional committees that were responsible for executing and implementing the project.

National Standing Committee

The national team was comprised of a total of nine (9) members.

- one National Project Manager
- two Lead Scorers (1 for the Arabic PT, 1 for the English PT)
- six Institutional Co-ordinators (each of the six participating institutions from the public and private sector nominated and appointed 1 Institutional Co-ordinator for the AHELO project)

Over the course of the AHELO feasibility study planning phase, the project parameters were continuously disseminated by the NPM to the ICs. ICs actively participated in sharing any issues and concerns that may have hampered planning efforts with the NPM, who brought them to the attention of the OECD during their planning meetings. During the planning stage of AHELO feasibility study, meetings were held every two to four weeks. Parallel to the meetings that were held by the National Standing Committee, an *ad hoc* Technical Committee was formed and responsible for the translation and cultural adaptation of the performance tasks. During the final five months prior to the student assessment, the planning committee met once per week.

National Technical Committee (ad hoc)

The technical committee was comprised of the NPM, one Deputy Manager, who also represented the Public Institutions, two ICs, who represented the Private Institutions, and translators who were nominated by the members of the National Team and vetted by the CM(s), according to the standards stipulated by the OECD/AHELO. Over the course of six weeks, the technical committee met at least twice a week to review the translations and cultural adaptations of the performance tasks. A pilot was conducted through the technical committee for the purpose of beta testing the translated performance tasks. It was the responsibility of the NPM to streamline the submission dates of the final documents with the OECD/AHELO.

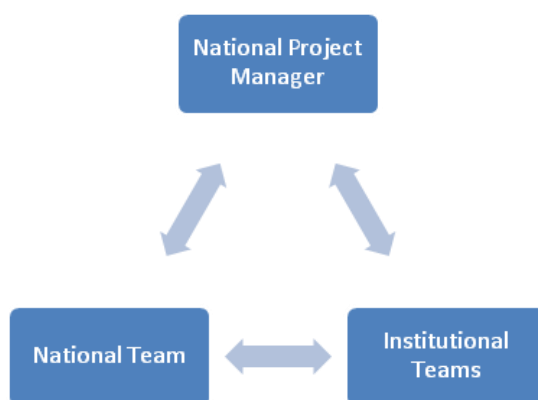
Institutional Committees

Each participating institution recognised the need to formulate a committee that would oversee the implementation phase of AHELO in each respective institution. Membership to these committees was determined by the IC in co-ordination with their respective academic and administrative authorities and required a minimum number of positions, as listed below:

- one Institutional Co-ordinator
- one Institutional Lead Scorer
- four Scorers
- two Technical Support Engineers
- administrative support personnel (flexible: determined by each institution as deemed appropriate)
- one Statistical Co-ordinator

Institutional Committees were chaired and managed by the Institutional Co-ordinator. The number of meetings was determined by the chair and each respective institution. During the testing day, additional staff volunteers provided support and assistance.

The management process with respect to the dissemination of information regarding project parameters, goals and outcomes adopted a linear, top-down approach. Information sharing and feedback adopted a more symbiotic relationship structure that encouraged open dialogue and creative problem solving. Please see the relationship flowchart below:



Preparation for fieldwork

During the planning and implementation phases of the AHELO feasibility study, the NPM encouraged and continuously supported the training of members of the national and institutional teams in the following areas:

Performance Task Academy

Sponsored by the Private University Council (PUC), the Performance Task Academy was offered by the Council for Aid to Education and conducted by Dr. Marc Chung. The objectives¹ of the academy were as follows:

- Participants were introduced to the basic architecture of the Collegiate Learning Assessment (CLA).
- Participants learned about the performance tasks, the use of rubrics, and overall aspects of scoring to assess higher order thinking skills like critical thinking.
- Participants went through the process of creating their own performance tasks that can be used as classroom tools.
- The Academy served as means of aligning teaching, learning and assessment.

Academy participation was managed through the ICs of each respective institution. Approximately three to four faculty members from each institution were selected in

accordance with institutional selection criteria. For the PUC, the purpose of sponsoring this workshop was twofold:

1. to provide familiarity with an example of an assessment measure that would be used in the AHELO; and
2. to encourage a cross-institution dialogue on assessment.

Translation workshop

Members of the National Committee attended the Translation workshop conducted by Dr. Solano-Flores during which he reviewed the eight steps of the translation process. The primary goal of the workshop was to provide assistance in the translation of the Performance Task that would support a cross-border feasibility study which would, in turn, ultimately determine the degree to which Performance Tasks could be translated and adapted to other languages and cultures.

Small scale lab validations (cognitive labs)

Participating institutions were requested to volunteer students from their respective institutions to partake in small scale validations. The cognitive lab was conducted by a member of the technical committee based on the criteria set by AHELO.

Training session and test administration

The Test Administration Manual was disseminated to all ICs followed by a workshop for Test Administrators and Scorers. The workshop was well attended: all members of the national committee were present, including most members of their respective institutional committees and teams. The workshop covered the basics on IT preparedness, test implementation and administration.

Individualised IT preparedness support

The National Committee provided continued support with respect to IT preparedness by scheduling on-site workshops at participating institutions, which included:

3. Discussions on security issues and concerns: security procedures were explained in the technical manual and discussed further with respect to disabling applications on all machines and enabling a compatible browser; ensuring that the enabled browser accessed the test site exclusively; Internet sites being blocked.
4. Checking system compatibility: a test was run on computer systems at each institution.

Sampling

Sampling of students and faculty were conducted in accordance to the AHELO criteria and by Institutional Research Offices.

Student samples ranged from 185 to 301 students. Students were randomly selected from registered students with 90 credits and above. Institutions presented their samples to the

National Committee with appropriate analysis in order check the degree to which the sampling captured a cross section of the student population regarding gender, nationality and major disciplines.

Faculty samples ranged from 15 to 60 faculty, depending on the size of the institution.

Fieldwork operations

Organisation and communication

The management structure and the inter- and intra-dependency of the committees that supported the development, planning, and implementation of the AHELO project, contributed to the timely and accurate flow of information. In case of miscommunication, the primary reasons can be attributed to a less than accurate understanding of assessment in general and the challenges that cross-border assessments pose, especially during a pilot on the part of a few participants.

Student recruitment: challenges and the question of incentive

From the initial planning phase of the AHELO, student recruitment was recognised as one of the more challenging aspects of the project, and was identified as a priority concern for the National Team. The National Committee identified several issues that require further deliberation and discussion:

- Making the assessment mandatory versus voluntary.
- The merits of incentive, monetary or otherwise.
- Credit based courses in which the AHELO generic skills test is an embedded assessment instrument within the courses.
- “Active” student participation: taking the test and being motivated to do well, versus “passive” participation: taking the test and rushing through it regardless of quality of performance.

Participating institutions were at liberty to develop their own incentive packages for students. Such incentive packages ranged from monetary packages to raffles in which students could win iPads. One institution decided not to provide an incentive package to students. Based on the results of student participation per institution, there existed no positive correlation between giving students “material” incentives to participate and the final participation rate.

Student active participation in extra-curricular educational activities that are taken seriously by students is of national concern. The notion that students only partake in educational events when there exists a measurable benefit, such as credit, needs to be further studied. The main question that must be addressed are the degrees to which students “do their best” on these assessments “when they don’t count,” and whether the results of the assessment truly reflect the student’s abilities when student attitude relative to the assessment is less than optimal.

Feedback from students/faculty***Student feedback***

Student feedback was mixed. Comments were related to the following areas:

- Degrees of difficulty and interest regarding the way in which the test was structured.
- Degrees of difficulty and interest in trying to understand the test content.
- Degrees of difficulty and interest in trying to use the test platform.
- Degrees of interest with respect to personal performance.
- Test length.
- Testing purpose: although students were provided with sufficient explanation as to the purpose of the test and the AHELO feasibility study, some students were still raising questions regarding the personal benefits of taking such a test.

Faculty/staff feedback

Faculty feedback was generally positive. There was a general consensus on comments related to the test platform in that it requires substantial enhancement. Faculty/administrators also felt that a student's non-familiarity with the structure of the test frustrated some students. Discussions related to whether students should be made familiar with the test structure prior to the test took place. Some faculty/administrators were uncertain the degree to which they should encourage "good performance" on the part of the students, as they were uncertain as to whether "motivational" factors were being measured in the test: in other words, they did not want to unduly interfere in the testing outcome with comments that may or may not motivate students. Again, the issues relative to "motivation" and measures of "motivation" need to be more carefully addressed in any future assessment plan.

Scoring process***Multi-language/multi-platform assessment***

The State of Kuwait chose to provide its educational institutions with the option of assessing its students' competencies in a language that best reflected its curriculum, its student cultural and academic profile, and in accordance with the predominant academic language that is operational at each respective institution/college.

In the case of students majoring in Quranic religious studies who took the written test over and above the computerised platform: the National Committee believed that students in more traditional majors were less comfortable with the computerised technology and word processing when applied to the Arabic language. The committee believed that the assessment of a student's generic skills should not be hampered by the student's inability or discomfort with the method of assessment and the platform used for assessment. The committee thus decided that providing those students with the option of a handwritten test eliminated a

variable that could potentially discriminate against students who are not sufficiently competent in Arabic word-processing because it is not a requirement in their field of study.

Scoring

During the implementation phase of the AHELO feasibility study, a total of 30 Scorers were identified. Each institutional committee selected one Lead Scorer, and four Scorers. Based on the final student participation numbers, the two Lead Scorers (one for the English language platform, one for the Arabic language platform) and five additional Scorers were selected for the final scoring. Scorers were vetted by the national committee, trained, and provided with a list of scoring criteria and rubrics.

The Arabic handwritten responses were scored by two Scorers. Extra security measures were taken in order to preserve the integrity of the grading process. Scorers were placed in the same venue and grading was conducted in the presence of the lead scorer.

Lead Scorers double-checked all scored tests.

Results

The national data files were distributed during the final week of January 2013 to participating institutions. Institutional reports are contingent upon the data files. Meetings have been planned for the National Committee to meet for the purpose of determining discussion criteria on respective institutional data results. Thus, institutional reports and discussions on information sharing remain forthcoming.

Impact at national/institutional/faculty level

National as well as international standardised assessments that provide comparative data are always useful and should be identified as one of many measures that determine the degree to which student learning outcomes are achieved.

The reliability of the AHELO Generic Skills strand as a cross-border instrument needs to be officially established prior to any other discussion regarding its impact on teaching and curriculum development. In the event that the AHELO generic skills test has been established as a reliable measure, the data is certainly valuable in terms of its comparative quality for local participating institutions with similar missions that are drawing students from the same population. This comparison will invariably encourage appropriate adjustments to teaching/curriculum development, and academic and institutional policies, if and when deemed appropriate.

On an international level, the data could measure the degree to which local institutions in any given nation measure over and against other institutions on broader global scale. This may extend the burden of responsibility of quality control beyond the institutional boundaries and may impact national policies regarding quality education, teaching and curriculum that include and go beyond secondary institutions of higher learning, and may very well involve public and private institutions from K-12.

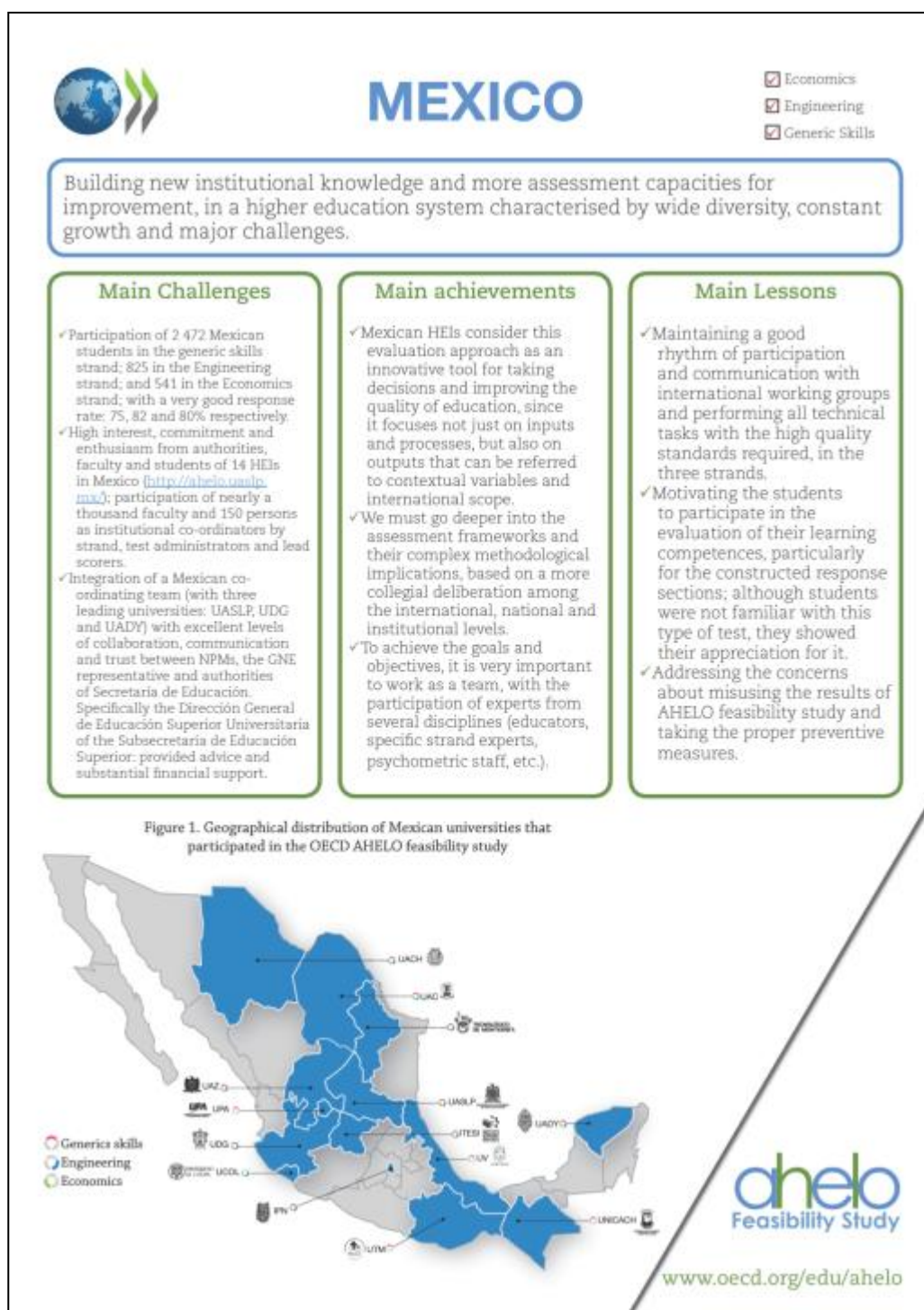
Three suggestions for a main study

- A study on student participation, incentive and test motivation.
- A study on the impact of cross-border assessment of secondary educational institutions on the National Education Policy of K-12.

Specific message

A study on the impact of cross-border assessment of secondary educational institutions on local accrediting agencies. A student, an institution, and a nation, capable of competing internationally, will be able to effectively partake in local, regional, and global economic progress and social development. Thus, the success of our students will positively impact the standard of living of everyone concerned.

Mexico



Rationale for the Mexican participation in AHELO feasibility study

Intense work has been done in Mexico during the past decades to build a quality assessment system, including self-assessment, external assessment, nationwide tests for entering and graduating students in higher education, quality assurance agencies, and the overall assessment of the system. It must be taken into account that the higher education system in Mexico is constantly growing, in the context of the demographic, economic, technological and political transition that the country is undergoing. Furthermore, a good part of the most influential Higher Education Institutions (HEIs) are implementing substantial reforms oriented towards innovation, faculty improvement, curricular flexibility, student mobility, and development of competencies, among others.

Mexico's participation in this project will allow for institutional knowledge on a new generation of assessment approaches, the exploration of methodological alternatives in the international spectrum, and the strengthening of the capacities of participating institutions and the entire HEI system. Participating Mexican HEIs in the AHELO project 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.

Participation of universities

The AHELO project in Mexico represents the high interest, commitment and enthusiasm from authorities, faculty and students of fourteen HEIs. It was possible to gather together a group of universities from different regions, and of different sizes, funding levels and academic degrees of autonomy.

The table below provides a summary of university types and participation in the three strands, and the figure below shows their geographical distribution. Because several universities participated in the three strands, it was necessary to appoint a university representative in addition to institutional co-ordinators by strand in each university, in addition to what was planned in the AHELO National Management Manual.

Participation of Mexican universities in the OECD AHELO feasibility study 2009-13

Universities	HEI type	AHELO Strand		
		Generic skills	Engineering	Economics
Universidad Autónoma de San Luis Potosí (UASLP)	State university (public, autonomous)	A	O	S
Universidad de Guadalajara (UDG)	State university (public, autonomous)	A	V	S
Universidad Autónoma de Yucatán (UADY)	State university (public, autonomous)	A	S	S
Tecnológico de Monterrey (Tec de Monterrey)	Private	O	O	O
Instituto Politécnico Nacional (IPN)	National university (public)	O	S	S
Instituto Tecnológico Superior de Irapuato (ITESI)	Technological Institute (public)	A	--	--
Universidad Autónoma de Chihuahua (UACH)	State university (public, autonomous)	A	S	S
Universidad Autónoma de Coahuila (UAC)	State university (public, autonomous)	A	A	A
Universidad Autónoma de Colima (UCOL)	State university (public, autonomous)	A	S	S
Universidad Autónoma de Zacatecas (UAZ)	State university (public, autonomous)	A	S	S
Universidad de Ciencias y Artes de Chiapas (UNICACH)	State university (public, autonomous)	A	--	--
Universidad Politécnica de Aguascalientes (UPA)	Polytechnic university (public)	A	--	--
Universidad Tecnológica de la Mixteca (UTM)	Technological university (public)	A	--	--
Universidad Veracruzana (UV)	State university (public, autonomous)	A	V	S
Total		13	10	10
Notes:	A = All campuses	O = One of several campuses	S = Single campus	V = Several campuses

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



National organisation

Mexico's participation in this feasibility study is a medium-term inter-institutional effort, under the leadership of three universities and the support of the *Secretaría de Educación*. A key factor was the organisation of a National Co-ordination Team with excellent levels of collaboration, communication and trust, composed of staff from the following institutions:

- The *Universidad Autónoma de San Luis Potosí*: responsible for the project's national co-ordination and for representing Mexico in the OECD Group of National Experts (GNE).
- The *Universidad de Guadalajara*: National Project Manager (NPM) and Lead Scorer of the generic skills and Economics strands, and responsible for participating in the Secretariat of the AHELO GNE.
- The *Universidad Autónoma de Yucatán*: NPM and Lead Scorer of the engineering strand.

- The *Secretaría de Educación* (SEP), specifically the *Dirección General de Educación Superior Universitaria* of the *Subsecretaría de Educación Superior*: provided advice and substantial financial support through the *Programa de Apoyo al Desarrollo de la Educación Superior*.

This team participated in the NPM and the OECD-GNE international meetings and maintained close communication with the OECD Secretariat, the Council for Aid to Education (CAE), and the ACER Consortium. In addition, the team received assistance from the Permanent Mission of Mexico to the OECD (*Secretaría de Relaciones Exteriores*) and the *Dirección General de Relaciones Internacionales* (SEP).

Project's stages: from design to fieldwork

All participants in Mexico's AHELO project actively contributed to all tasks required in the design and implementation phases: development of assessment frameworks; instrument design, translation, small-scale validation and cultural adaptation; HEI invitation, selection and organisation; student and faculty sampling; technical preparation; test administration; scoring training, calibration and capture; and preliminary data analysis.

To support these activities, the team kept close communication by the usual means and conducted national meetings and workshops in order to establish necessary agreements, participants' training, and resolve questions and concerns. The table below shows the main national events that were organised.

Main national events organised in Mexico's AHELO project

Event	Place and date	Participants
National Workshop: Background, objectives and progress of the AHELO project	San Luis Potosí, SLP; 13 December 2010	National Co-ordination Team and university representatives
National Workshop: "Performance Task Academy" by Council for Aid to Education	Guadalajara, Jal; 20-21 December 2010	Generic skills strand: NPM, institutional co-ordinators and some HEI faculty
National Meeting: Progress of the AHELO project and prospects for the implementation phase	Mexico City; 9 November 2011	National Co-ordination Team, university representatives and institutional co-ordinators by strand
National Workshop: Students and faculty sampling, technical preparation, test administration, and information management criteria	Mexico City; 31 January and 1 February 2012	National Co-ordination Team, code leaders, and institutional co-ordinators by strand
National Workshop: Students and faculty recruitment, electronic system test and contextual dimension	Mérida, Yucatán; 27-28 February 2012	National Co-ordination Team, university representatives and institutional co-ordinators by strand
National Workshop: Test administration, quality monitoring, national scoring team and scoring operations	Guadalajara, Jal; 26-27 March 2012	Generic skills strand: NPM, lead scorer, institutional co-ordinators and some potential test administrators and scorers
National Workshop: Rubrics, calibration and scoring processes	Guadalajara, Jal; 28-30 April 2012	Generic skills strand: NPM, lead scorer and scorers.
National Workshop: Assessment of Higher Education Learning Outcomes (international trends, competences conceptualization, task performance design)	Mérida, Yucatán; 6-8 December 2012	National Co-ordination Team, university representatives, institutional co-ordinators by strand and scorers
National Meeting: Results of AHELO implementation phase, analysis of the structure of expected reports (international, national and institutional); and future scenarios	Mérida, Yucatán; 8-9 December 2012	

Additionally, between 2009 and 2012, the National Co-ordination Team held fifteen face-to-face meetings in the cities of México City; San Luis Potosí, SLP; Mérida, Yucatan and Guadalajara, Jalisco.

In order to complement the co-operation between HEIs, stakeholders and public, the Mexico AHELO project's website (<http://ahelo.uaslp.mx>) was continuously updated, including basic information, web links and downloadable resources. Furthermore, promotional materials were designed and printed for all stages of the AHELO feasibility study. Specifically for the implementation phase, posters, postcards and brochures were distributed for students, teachers and the public in general.

Photo of the National Co-ordination Team, university representatives, institutional co-ordinators by strand, and some scorers; Mérida, Yucatan, National Meeting, December 2012



Goals and results

Mexico's participation in the feasibility study will be completed in 2013, with the publication of the reports and the organisation of a conference at the international level. Between February and May, data will be analysed. In June 2013, a workshop with all participating HEIs is expected to be held to analyse the impact of the AHELO project, as well future actions. Specifically, Mexico will also produce a national report to be published at the beginning of 2014. The table below shows the main goals achieved until January of 2013.

Main goals achieved up to January 2013 for Mexico's participation in the OECD AHELO feasibility study, 2009-13

Goal	AHELO Strand		
	Generic skills	Engineering	Economics
Students sampled	2 472	825	541
Students tested	1 842	678	402
Response rate	75%	82%	80%
Faculty	400	366	217
Institutional coordinators	13	10	10
Test administrators	56	29	23
Test scorers	14	9	11
Sessions	68	23	21
NPMs, GNE and staff support	8	3	4

Students found the AHELO project interesting, but challenging, since they were not familiar with this type of test. Moreover, they had difficulties during the test because they had forgotten some issues studied early in their career.

Challenges

The main challenges experienced in Mexico were:

- 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 in which Mexico decided to participate: Generics, Engineering and Economics.
- Motivating the students to participate in the evaluation of their learning competences, particularly for the constructed response sections. Students participated very enthusiastically in the AHELO test; however they are not familiar with this type of test and some of them thought that, considering the length of the test, the time given to complete it was insufficient.
- Addressing the concerns about misusing the results of the AHELO feasibility study, and taking the proper preventive measures.

Suggestions for a main study

For a main study, we have the following suggestions:

- We must go deeper into the assessment frameworks and its complex conceptual and methodological implications, based on a collegial deliberation among the international, national and institutional scopes.
- 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.), in all levels of study (international, national and institutional).
- It would be helpful to define more clearly the benefits of the project for all participants, i.e. for policy makers, institutions, teachers and students.

Netherlands



NETHERLANDS

- ☒ 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

Key data

The Netherlands have a binary system of Higher Education, with Research Universities (*universiteiten*) and Universities of Applied Sciences (*hogescholen*) offering programmes covering the entire academic field.

All Research universities offer programmes traditionally more research and knowledge-based, some also offer programmes more practically or applied oriented (for instance Business Schools), Universities of Applied Sciences offer programmes that are practically oriented.

Research Universities offer all three cycles: Bachelor, Master and PhD. Universities of Applied Sciences offer mainly Bachelor programmes and some Master programmes.

For the Economics discipline the Dutch government offers funding to public universities, for all Bachelor programmes, for Master programmes at Research Universities, and for PhD programmes.

The Netherlands decided to participate in the AHELO feasibility study in the Economics Strand. This decision was taken by the Dutch Ministry of Education, Culture and Science, the National Association of Universities (VSNU) and the National Association of Universities of Applied Sciences (HBO-raad). The Universities' Boards were informed on the feasibility study, and asked to express their interest.

Initially, thirteen Higher Education institutions expressed their interest in participating in the AHELO feasibility study. This included three research universities, nine universities of applied sciences and one private university. Several informative sessions were organised by the Dutch Ministry of Education, Culture and Science and a National Programme Manager was appointed in 2011.

The long-term uncertainty of the project's funding status made it difficult to ask the institutions to become actively involved, with the situation being unclear if field work would actually be possible to organise. This lasted until late 2011, so only in early 2012 could the interested institutions be invited to a session to explain into detail what participation in the feasibility study would entail.

The implementation of the feasibility study would have to take place within the following five months, with a deadline for the student assessments and scoring of the results by the end of June 2012. For most institutions this timeline was too limited to get their active participation organised, especially with all other quality measures already being implemented and taking up the available time and attention (such as accreditation processes). One university explained not being interested in participating in the study in a feasibility stage, others explained they could not free up staff and time to participate.

In March/April 2012 it was clear that five institutions could participate in the feasibility study: one research university and four universities for applied sciences. Within these institutions, 109 faculty members were selected through a random selection process and asked to fill in the faculty context instrument, and approximately 600 students were invited to participate in the assessment.

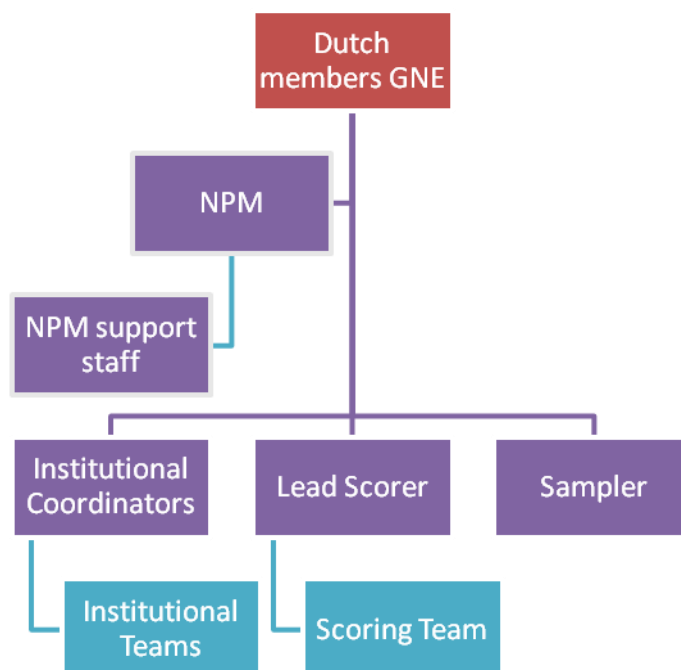
Total faculty response: 31; total student response: 20.

National and institutional management

The Netherlands decided not to set up a full-scale National Centre for the co-ordination of the feasibility study, with the small scale of the country geographically and the small number of institutions participating in the field work. Management of the field work was done by the National Programme Manager (NPM) and support staff, with the Dutch representatives in the Group of National Experts acting as a steering group but also as very active ambassadors of AHELO.

Each institution provided an Institutional Co-ordinator: a staff member on institutional level who was the main contact person within the institution and for the NPM. For the actual field work one or more staff members from the Economics department or school were added to the institutional organisation.

The Lead Scorer was responsible, with the NPM, for setting up a National Scoring Team, which in the Netherlands – with the low response rate – was a very small team, consisting of two members. These were experts in the field of Economics. Another team member was added for the sampling work, with expertise on SPSS.



Preparation for field work & field work operations

In preparing for field work, all initially interested institutions were approached following the kick-off meeting at the Dutch Ministry in January 2012. Most institutions were visited by the NPM, some several times. When the final 'go' was given at the participating five institutions, a variety of actions was undertaken to support them: two institutions managed to organise the field work themselves, three others asked the NPM to take up the local organisation for them.

For these last three institutions the NPM added some temporary staff to the support staff to assist in organising facilities (including testing of the equipment needed), sending out invitations to and answering questions by students and faculty, promoting AHELO on-site, and for the test administration.

For all institutions a Dutch translation of the Institutional Co-ordination Manual, the Test Administration Manual, summaries of both documents and the OECD AHELO brochure was provided.

A flyer describing the assessment was produced in Dutch for students and staff, as well as a website. Posters were produced and distributed to the participating institutions with a reference to the website to increase awareness amongst students and staff.



For the two institutions that organised the field work themselves, several meetings with the Institutional Co-ordinators were scheduled, as well as a training for test administration (in one institution as a train-the-trainers session). For the other two institutions test administrators received training in a central location.

The Lead Scorer attended meetings organised by the Consortium in Paris, to discuss and specify the Scoring Manual. A session was organised in the Netherlands to discuss and train the scoring process, and the actual scoring was done on one day.

The Dutch Ministry of Education, Culture and Science provided the opportunity to offer an I-pad for each institution to be put up for raffle amongst the students participating in the assessment. Two institutions incorporated the assessment in a session with the final year students to evaluate their experiences with the programmes and the institution, and offered respectively a dinner beforehand and a drink afterwards. One institution sent out invitations to students and staff by the University Board, one institution by the Dean of the Economics

Department, one institution by an internally well-known staff member, and two institutions by the NPM. Both students and staff received reminders in all institutions.

With the small numbers participating in the assessment, especially students, it is of course difficult to generalise, but participation appeared to be more appealing to students if the invitation was sent by someone familiar to them (Dean or staff member) and if the assessment was incorporated in a more extensive session. Eventually participation was very limited, because of the compressed timeline and because of the suboptimal timing of the assessment within the faculty's academic year.

Feedback

Feedback from the Institutional Co-ordinators showed that the materials provided were clear, and offering helpful guidelines in organising the field work. The documents provided were sometimes seen as too extensive (this goes especially for the Test Administration Manual), so the summaries also provided were considered convenient.

The Institutional Co-ordinators were able to gather feedback from the students, from those who did and who did not participate in the assessment. The students who did not participate in general gave two reasons for this: lack of time and the fact that they did not see what was in it for them to participate in the assessment.

The small-scale validation of the assessment that took place at an earlier stage showed a (much) higher participation rate, and although in general the invitation came from a closer contact of the students, the fact that a financial compensation was being offered should also be taken into consideration (and was in fact also part of the feedback students gave then).

The lack of a clear incentive for each individual student, either a financial compensation or at least individual feedback on the results of the assessment, was mentioned as a reason not to participate.

The students who did participate in the assessment gave feedback on both the set-up of the assessment and the contents of the test. The set-up of the assessment was considered to be good. In two institutions there were some minor technical problems that were solved quickly. There was no difficulty in the test being administered online. Also the length of the assessment was considered to be alright.

As far as the contents of the assessment were considered², in all institutions (both research universities and universities for applied sciences) students mentioned the assessment being more knowledge-based and less practical oriented or competence-based than they were expecting. Students in both types of participating institutions said that the knowledge being tested in the assessment had been part of their study programme to a large extent, but also that the subjects covered in the assessment had been studied in an early stage of their programmes. They did not see recent course work reflected in the assessment.

In the feedback students gave to the Institutional Co-ordinators, there was a distinction between students from a research university and those from a university of applied sciences.

The latter expressed much more uncertainty about their performance on the test. Students from the participating research university seemed more confident about their performance.

This may reflect the concern raised by the Netherlands in several meetings, in GNE, NPM and Lead Scorers' sessions, that the aim to offer a single assessment in a binary Higher Education system might not be catering for all needs.

Impact


It may be clear that from a feasibility study with such unfortunate low student participation numbers, the impact is very limited, on all levels (national, institutional and faculty level). Based on N=20 it is actually not possible to draw any conclusions for the Dutch situation for the content of the survey or for the practical feasibility of an Ahelo. However we can draw some very tentative conclusions. The feeling is that the items in the economics strand are more knowledge-based than would be expected and they do not cover the whole range of economic studies provided in the initial economics framework. Seeing all the effort at faculty, institutional and national level, the practical feasibility doesn't give much hope for a full fledged assessment.

Suggestions for a main study

For a main study on the Assessment of Higher Education Learning Outcomes the following suggestions are being made:

- In the development of future assessment instruments 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. This goes especially for the development and selection of items within the disciplinary framework.
- Try to incorporate the instruments more into the quality assurance measures already being implemented in Higher Education on a national and institutional level. This way institutions will be able to create facilities needed to organise field work and evaluate the outcome as part of processes already taking place.
- Try to incorporate 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.
- 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.

Norway



NORWAY

☐ Economics
☐ Engineering
☒ Generic Skills

The study offered a chance to learn more about the outcomes of higher education and the potential use of the generic skills tests. It showed us that engaging students and staff is a central challenge in such work.

Main Challenges


- ✓ Response rate: the most significant challenge by far was recruiting enough students, and the response rate was lower than hoped for.
- ✓ High demands on institutional resources: the complex scoring, substantial work to recruit for and run the survey and the need for large student incentives (prizes/giftcards) raised costs.
- ✓ Timing: the test period was in a semester when many students spend less time on campus and instead work on projects or a bachelor's thesis.


Main achievements

- ✓ Good co-operation within the Norwegian team: the planning, technical preparation and collaboration between the institutions and national team worked well.
- ✓ It showed that it is feasible to deliver electronic tests globally: aside from a handful of minor technical problems, the testing process ran smoothly.
- ✓ Trying out a variety of approaches to promotion and incentives: institutions tried a range of approaches to inform and attract students to this pilot.

Main Lessons

- ✓ The processes and collaboration required for such tests are in place, but the balance of costs and benefits for students and institutions needs to be considered carefully in the future.
- ✓ The institutions are interested in such tests, but the time and resources demanded in this case were high, especially considering the limited information offered by results (due to response rate).
- ✓ There is no obvious solution as to how students can be encouraged to take such an extensive test – a more targeted sampling approach may prove more practical and effective in Norway.




www.oecd.org/edu/ahelo

Key data on participation

Norway took part in the Generic Skills strand only. The test was administered at five Norwegian institutions. Following the sampling approach set out by the consortium, Norway approached initial samples of 1 500 students (300 from each institution) and 300 faculty members (60 from each institution).

National and institutional management

The Ministry of Education had formal responsibility for the AHELO feasibility study in Norway. The national project management and task of carrying out AHELO and co-ordinating co-operation between the main AHELO consortium, the Ministry of Education and the participating Norwegian institutions, was put out to competitive tender. The Nordic Institute for Studies in Innovation, Research and Education (NIFU) was awarded this role, along with EKVA (Unit for Quantitative Analysis in Education) at the University of Oslo as a subcontracted expert to handle translation and scoring of constructed response tasks.

The Ministry of Education recruited Higher Education Institutions to participate in the study. The Norwegian University of Science and Technology (NTNU), the Norwegian University of Life Sciences (UMB) and Vestfold University College (HiVe) agreed to participate right from the start of the study. In 2011, two further institutions joined: the University of Stavanger (UiS) and Lillehammer University College (HiL). These five participating institutions were quite different from each other, both with regard to size and mission, type of study programmes and student composition.

The National Project Manager (NPM: NIFU), the Ministry and these five institutions met regularly and worked closely together throughout the project phases; the key team members from EKVA/ILS were heavily involved during the translation, scoring training and scoring phases.

Preparation for fieldwork

During the preparation phase, the key tasks were: translating, adaptation and checking the test instruments, scoring rubrics and other documents and manuals; gathering and preparing the samples for students and faculty; and, putting practical plans and technical tests in place in institutions. There was also substantial work preparing and training the scoring team (see below).

Translation and adaptation

The process of translating and adapting the training material was done via two separate processes: the translation of the constructed response task (CRT) was done in 2010, while the translation of multiple choice questions (MCQs) and context instruments started after the decision to implement the study had been reached (July 2011). Thus, these instruments had to be completed under greater time pressure. It was important and useful to have started the CRT translations early, as these required more substantial work (due to the lengthy document libraries) while the MCQ translation process went quite quickly and smoothly, supported by good routines for feedback and review between the central consortium and national team.

Translating the scoring instrument/rubric for the CRT was also a challenge, as it included nuances in English that could not be directly translated into Norwegian. In the process of translation, the potential issues regarding consistent interpretation, meaning and maintaining similarly challenging tests in different translations were apparent, but these were generally resolved in a way that all those involved felt confident about. Overall, the translation and adaptation process seemed to work well, but were a little more resource-intensive than anticipated.

Sampling

The sampling process for students and academics went smoothly in terms of gathering necessary data from the participating institutions and following the technical steps. As the AHELO feasibility study did not seek to compare at the higher education system level, institutions were simply selected via convenience sampling – based on those willing and enthusiastic to take part. Students were randomly sampled across disciplines and programmes at each institution (300 students per institution) based on records at the participating institutions, which were of a high quality (complete) and relatively easy to access. Each institution's sample was checked as a broad fit with the disciplinary composition of the institution, and all of them were a good match. The faculty (staff) were also randomly sampled across all disciplines (60 per institution); however there was no connection between the faculty sample and the student sample, as the scoring manual did not state that as an ambition of the feasibility study.

The sampling manual provided for the AHELO project was detailed and the processes were followed without any problem. The team did discuss the pros and cons of the selected approach during this phase; while a random sampling approach is rigorous, a more targeted selection of students (e.g. stratified sampling within a sub-set of programmes) might have supported more targeted recruitment strategies, and the chosen design did not provide a connection between the student and the faculty (no matching of selected students and faculty), meaning faculty statements might only provide a limited clarification regarding the student responses.

Fieldwork operations

Norway started the AHELO test phase in early February 2012. Thus, Norway was the first country to start testing. The final preparations and technical testing had gone well and proved valuable as a way of identifying any challenges in the IT systems at each institution. For example, setting up a secure test environment (blocking access to other websites during the test period) was challenging in some cases, but solutions were found to ensure this functioned well during the actual test. Very few technical problems occurred during the test phase, apart from a small number of tests that did not run to completion.

The overall organisation of the testing phases went well: the institutions and National Project Manager were in regular contact and worked together to resolve technical challenges and address questions. However, one week of testing had to be cancelled at the institution that was to start first, due to delayed delivery of log-in codes, and this probably caused the loss of a few participants. Institutions had started recruiting promptly, asking students to sign up in advance

for testing sessions. It was also evident early on in the testing process that recruiting enough students to take the test would be extremely challenging, and this was the main focus of additional effort and activity.

Students had been recruited using a range of approaches. All sampled students received an email invitation to take part and could sign up for a testing session of their choice. Institutions also used a range of measures to inform students about AHELO, such as articles in student newspapers and information on homepages. Institutional Co-ordinators worked hard to provide information and encouragement in the weeks before testing with recruitment and follow up on the students. All institutions also offered different forms of incentives to students upon completion of the AHELO test, either monetary rewards/gift cards or lottery.

Feedback from students/faculty

Some feedback was provided via institutional contacts, regarding participants' views. Students expressed a reluctance to take part in such a lengthy test that did not provide them with a grade or a credit. The views of those who did take part varied from finding it interesting, to boring and too long.

Scoring process

Due to the low number of responses, the scoring of all CRTs was conducted by two Scorers, trained and overseen by the Lead Scorer. All responses were scored twice, and in cases with a substantial deviation in scores, the Lead Scorer also scored that response and made a final decision. Scoring took place in the first two weeks of May and went smoothly. However, feedback from the Lead Scorers indicate that having a more tailored scoring rubric for each of the tasks, instead of a general scoring rubric, would have been helpful, and that it was hard to distinguish between "borderline" responses in some cases.

Results

The response rate among students in Norway was disappointing: the national average response rate was just 7.7%, with individual institution's rates ranging from 4.7% to 10%.

Key reasons for the low response rates are thought to be the long duration of the Generic Skills test (over two hours) and the timing of the test period; many students were in a period of their course where they are expected to work on a thesis project, with few structured classes and little time on campus.

The institutions all provided feedback on their experiences with AHELO to the NPM. They were generally positive about the organisation of the project, communication within the national team and the technical processes involved. The areas that were more often a cause of concern were recruitment of students to participate in the study, which had proved resource intensive and challenging. Institutions also reported higher than expected costs of scoring.

Impact at national/institutional/faculty level

The study was a source of great interest, as it offered a chance to learn more about the outcomes of higher education and the potential use of generic skills tests. It showed that it is

feasible to deliver electronic tests globally, as aside from a handful of minor technical problems, the testing process ran smoothly.

It is too early to identify most potential impacts, as the reporting process and consideration of the data is underway; furthermore, the low response rates mean the Norwegian data offers only limited analysis and does not provide a robust basis for comparing between sub-groups or institutions.

Attempts to measure learning outcomes directly are still of great interest to Norwegian actors and higher education institutions, in light of the considerable limitations to self-assessments/ratings of skills (e.g. relativistic answers). However, in order to be able to measure learning outcomes directly, it is essential to find a way to engage students in such studies.

Any particular innovative process you would like to share

As the challenge in recruiting students became clear, the national team and institutions worked hard to identify and try out a variety of approaches to promotion and incentives; this was a valuable aspect of the pilot. None of these proved decisive, suggesting that financial incentives/prizes are not enough to motivate most students in Norway and closer consideration of how to engage them in future studies is needed.

Any particular challenge or problem you met?

The most significant challenge was recruiting students, and response rates were far lower than hoped. Exam periods and independent study times were a challenge to recruitment of students in Norway. Response rates may also reflect relationships between Higher Education Institutions and their students to some extent; it may be that being chosen to participate in such an international study is seen as an honour or an obligation in some countries, but Norwegian students did not seem to consider participation in such terms.

AHELO has received considerable interest from governments and institutions and, if anything, this interest has increased over time. However, it seems as if one of the most important stakeholders, the students, need to be further engaged and motivated to volunteer their time to participate.

Three suggestions for a main study

- While few technical problems were encountered, the testing phase was essential and helped to avoid potential problems. Maintaining this technical testing and “dry run” of tests is vital, even where online test tools are well-trialled, as individual institutions faced different challenges regarding their IT infrastructure.
- There is considerable interest from institutions in such tests, but even with good and effective international co-operation, nations will likely face diverse challenges regarding practical issues, such as course structures and term times, and cultural differences, such as attitudes to testing. These issues should be considered in the test design.

- The processes and collaboration required for such a test are in place, but the balance of costs and benefits for students and institutions needs to be considered carefully in future. In particular, more effort will need to be made in identifying how students can be interested in participating in much higher numbers.

A specific message from Norway

“It’s like sitting an extra exam – but what do you get from it?”

A key message from the Norwegian experience, which may require greater consideration in future studies of students’ learning outcomes, is the need for a clearer narrative about what such processes can offer to students taking part. Such testing needs substantial input from many participants, and this does not seem to be motivated by incentives alone. A clearer sense of how the test results, and the overall study, may benefit the participants (both students and institutions), as well as how it may contribute to further developing the quality of HE, might be an important issue that could support higher response rates and future work in this area.

Russian Federation



RUSSIAN FEDERATION

☒ Economics
☒ Engineering
☐ Generic Skills

In the Russian Federation, the AHELO feasibility study is a large-scale bottom-up project, initiated and coordinated by universities (Higher School of Economics in cooperation with the Ural Federal University in Engineering) and accomplished with great support and a high level of participation of Russian universities, serving as a basis for the further exerts networking and cooperation in QA.

Main Challenges	Main achievements	Main Lessons
<ul style="list-style-type: none"> ✓ The nature of tasks for testing: theoretical tasks prevailed, the limited bank of tasks and a high risk of cheating. ✓ The legal and practical difficulties of collecting student and faculty personal data for sampling. ✓ The differences of curriculum between HEIs and difficulties for further comparability. 	<ul style="list-style-type: none"> ✓ Large scale of university participation, involvement of regional universities in this international research. High level of support from students, faculties, universities administration. ✓ High response rate among students. ✓ Elaboration of supplementary national instruments (online questionnaires for students and faculties) to get additional data about LO context. 	<ul style="list-style-type: none"> ✓ Independent national supervisors are needed to assure compliance with the testing procedures and minimize the risk of cheating. ✓ The bank of tasks needs to be larger and more diversified, it would be better to link the tasks with the real professional activities. ✓ The HEIs are able to organize the assessment process following the international guidelines but national project management system is needed to assure their coordination and communication.

Key message: The AHELO project is in line with the national strategic objectives in the QA area relating to the creation of the independent quality assessment system in higher education and encourages such current national initiatives as a Federal exam for bachelors (since 2012). The AHELO feasibility study was the opportunity to have an experience of measuring higher education competencies on the national and international level, the possibility of comparison of educational quality level in different HEIs and possibility for federal and regional universities to participate in international research on the same level and the chance to bring students learning outcomes of different universities to the one scale.





www.oecd.org/edu/ahelo

Key data on participation

Russia participated in 2 strands: Economics and Engineering

In total 26 universities participated, including:

- Three universities in both strands
- Seven universities in the Engineering strand only
- Sixteen universities –in the Economics strand only (including two in Phase 1 only)

All Russian regions and different types of universities participated in the AHELO implementation, including:

- Six Federal Universities
- Four National Research Universities
- Two main first-level multi-profile universities (St. Petersburg and Moscow State Universities)
- Five polytechnic universities
- Six universities specialised in Economics, Public Administration or Management
- One classic multi-profile university

In total 3 581 students and 520 faculty took part in the field research. More than 300 people were involved into the project implementation (ICs, TAs and TAs' Assistants, national experts, interpreters, etc.)

	Target population size	Sample size	Participation size	Response rate
Students - Economics	3756	2900	2400	82%
Faculties - Economics	1556	714	349	49%
Students - Engineering	1282	1282	1181	92%
Faculties - Engineering	331	331	171	52%

National and institutional management

The project in Economics is coordinated by the National Research University – Higher School of Economics (HSE). The project in Engineering is coordinated by the HSE in cooperation with the Higher School of Engineering of the Ural Federal University named after the First President of Russia B.N. Yeltsin (UrFU).

The National Coordination Centre is established at the National Research University – Higher School of Economics with the following composition of the team:

- Mr Isak Froumin, Head of the AHELO National Expert Council (HSE)
- Ms Tatiana Meshkova, NPM in Economics and co-NPM in Engineering (HSE)
- Mr Oleg Rebrin, co-NPM in Engineering (UrFU)
- Ms Elena Sabelnikova, Assistant to NPM in Economics and Engineering (HSE)
- Ms Irina Sholina, Assistant to NPM in Engineering (UrFU)
- Mr Vladimir Zuev, Member of the AHELO Economics Expert Committee (HSE)
- Mr Kirill Bukin, Member of the AHELO GNE, Institutional coordinator at HSE
- Ms Veronika Belousova, Lead Scorer in Economics (HSE)
- Mr Vladimir Volkovich, Lead Scorer in Engineering (UrFU)

The National Expert Council including experts from academic and Ministry's community was created.

Scorers team:

- Economics: Lead Scorer and seven Scorers
- Engineering: Lead Scorer and eight Scorers

Institutional coordinators: 19 in Economics and 10 in Engineering

Experts networks in QA in Economics and Engineering were formed

National web-site: www.hse.ru/ahelo/

Preparation for fieldwork

Training

- Training-webinar for Institutional Coordinators for sampling procedures discussion: 27 and 30 January 2012 (through Webex)
- Training-webinar for ICs and TAs to discuss the field study schedule, preparation and implementation: 22 and 26 March 2012 (through Webex)
- Training for scorers: April 2012

Sampling

- Economics: One-stage simple designs
- Engineering: total evaluation

To determine the general population a special form for data collection according the target groups of students and faculty staff was developed.

Due to that fact that in some HEIs the transition from specialist to two-tier system of education (bachelor, master degrees, 4+2 years) is continuing we came across the problem of lack of bachelor students for the study. So we had to include 4-year specialists along with the bachelor degree students.

Fieldwork operations

Schedule of testing

	HEIs	Dates of testing
Both strands		
	Northern (Arctic) Federal University	23- 24 April 2012 (Economics) 21 May 2012 (Engineering)
	North-Eastern Federal University named after M.K.Ammosov	10, 11, 12, 13 April 2012 (Economics)
	Ural Federal University named after the First President of Russia B.N. Yeltsin	11 April 2012 (Economics) 2-4, 10-11, 21-22 May 2012 (Engineering)
Economics		
	The Russian Presidential Academy of National Economy and Public Administration	4 April 2012
	The State University of Management	23 March 2012
	Far Eastern Federal University	10, 11, 12 April 2012
	Kazan (Volga Region) Federal University	27 April 2012
	Moscow State University of International Relations	18 April 2012
	Moscow State University of Economics, Statistics and Informatics	19 April 2012
	National Research University Higher School of Economics (Moscow)	11, 17, 18 April 2012
	National Research University Higher School of Economics (St. Petersburg)	17 May 2012
	National Research University Higher School of Economics (Perm)	11, 12 May 2012
	National Research University Higher School of Economics (Nizhni Novgorod)	10 April 2012
	National Research Novosibirsk State University	3 April 2012
	Plekhanov Russian University of Economics	2, 3, 5 April 2012
	St. Petersburg State University	26, 27 April 2012
	St. Petersburg State University of Economics and Finance	16, 17 May 2012
	Siberian Federal University	21 April \ 2012
	Southern Federal University (Rostov-on-Don)	19 April 2012
	Southern Federal University (Novoshakhtinsk)	23, 24 April 2012
	Southern Federal University (Taganrog)	18 April 2012
	Altai State University	14, 15, 16 May 2012
Engineering		
	National Research Irkutsk State Technical University	23, 24 May 2012
	National Research Tomsk Polytechnic University	4, 5 May 2012

	Tyumen State Oil and Gas University	15-21 May 2012
	Ural State Mining University	17 May 2012
	Ural State University of Railway Transport	22 May 2012
	Don State Technical University	11, 12 May 2012
	I.I. Polzunov Altai State Technical University	17, 18 May 2012

Student recruitment: For the purpose of engaging students the HEIs used administration measures as well as different types of motivation. For example, in some HEIs student participation in AHELO was awarded to the practice or was rewarded with different small presents and bonuses. But the main instrument for motivation was the international OECD certificate of participation.

Feedback from students/faculty

Feedback on assessment instruments:

Tasks:

- The majority of items were standard tasks from the books (especially in Economics). Constructed response tasks were more interesting.
- The test was difficult for students because they rarely come across such format of the items.
- Students are not taught all the themes included in the test within their curriculum.

Questionnaires:

- Questionnaires (especially for the Institutional Co-ordinators) were not exactly adapted to the peculiarities of Russian educational system. It was difficult to answer some of the questions.

Feedback on process

- The online platform is much more comfortable than paper-and-pencil.
- Technical support was needed because of problems with test server mistakes and errors.

Feedback on test length

Test and questionnaire length as a whole was acceptable for the majority of the students.

Scoring process

Economics: The process was organized due to the OECD requirements. Some problems took place. The Lead Scorer did not see the other scorers' marks and could not change anything as a Leader. So they wasted their time to find the divergence of their opinions. Also the system often hung.

Results

The main result is a high interest and support of Russian HEIs which reflect the profound changes in QA, management and internationalization processes in Russian HE.

The feasibility of the AHELO methodology is proved in Russian HEIs: two phases of the project have been fully implemented and the national assessment instruments (national questionnaires) have been used to get additional information about the context surrounding learning outcomes

This project had an influence on reforming the contents of Russian education. New state educational standards are based on the results of education and take into consideration the AHELO methodological approaches.

Response rates achieved is indicated above

The total number of Institutional Co-ordinators reports received was very small (2) because the timing to answer the report coincided with the beginning of summer leave. Also the reports were to be done in English but the funding for that was not provided in the national budget of the project.

Impact at national/institutional/faculty level

- At institutional level: reflection on curriculum (serious cross-institutional differences of curriculum), networking and discussion on assessment instruments and process.
- At national level: the AHELO experience was used to implement a Federal exam for bachelors (a pilot national project started in 2012): the further national researches taking into account the AHELO experience are included to the National programme "Development of Education from 2013 till 2020".
- Taking into account the project methodology and the AHELO methods of assessment of learning outcomes some changes to the design and realization of bachelor educational programmes were made in some HEIs (Engineering).

Any particular innovative process you would like to share

We would like to share our experience to use some supplement national assessment instruments along side with the international instruments (online questionnaires for faculties and students) to get additional context data for national analysis.

New approaches for degrees classification in Engineering based on the OECD FIELD OF SCIENCE AND TECHNOLOGY CLASSIFICATION are proposed to HEIs and to the Ministry for consideration.

Any particular challenge or problem you met

- Modest level of faculty participation. Additional motivation and organisation measures need to be taken to assure a sufficient participation rate among faculties.

- Scale of universities participation vs. effective project management: large territory, time difference, large sample, etc.
- Nature of tasks: the constructed response tasks were more unusual and difficult but more interesting for Russian students.
- In some universities the students met with several instances of online system errors and interruption during test sessions.
- Not all of the Engineering HEIs which were interested in AHELO could take part in the field study because the official decision on the RF participation in Engineering strand had been taken only at the end of April 2012, while the final university exams in Engineering HEIs began in May. Consequently the time of testing and the time of the final exams were the same.
- There is no degree programme of Civil Engineering in Russian HEIs. It was necessary to run the preliminary comparative analysis of international and Russian classification coding systems for the purpose of defining Engineering degrees delivered by Russian HEIs which would be able to include in the Civil Engineering.

Three suggestions for a main study

- The bank of tasks needs to be larger and more diversified.
- The online test system needs to be more secure.
- The access to the data (national and international) for national centres needs to be more complete and prompt.

Slovak Republic



SLOVAK REPUBLIC

☒ Economics
☒ Engineering
☒ Generic Skills

For Slovakia, the AHELO feasibility study represented a light at the end of the tunnel in the endless debates on the quality of Slovak higher education.

Main Challenges	Main achievements	Main Lessons
<ul style="list-style-type: none"> ✓ Finding ways to motivate HEIs and students to participate in the AHELO project. ✓ Managing, in a short time and without any previous experience of this type, a project of such size and complexity in the HE environment. ✓ Taking the risk of a relatively high financial investment in a project for which the development and results were uncertain at the moment of making the decision. 	<ul style="list-style-type: none"> ✓ High participation of HEIs and high, in some cases even very high, participation of students in testing. ✓ Obtaining new information on the Slovak HEIs based on methodology guaranteed by and international consortium. ✓ Successful organisational and technical management, including good decisions regarding the technical arrangements (mobile computer labs), leading to smooth implementation. 	<ul style="list-style-type: none"> ✓ It is possible to motivate the HEIs and students to take part in innovative activities and to engage them with the offer of obtaining information on their results in order to compare them with those of their peers. ✓ Success in a project of this kind requires the synergy of several factors: enthusiasm and professionalism of people; adequate funding; willingness to take risks; and the support of the Centre. ✓ In principle it is feasible to get external and internal comparisons of HEIs of different countries.

Key message: For a small country like Slovakia, the use of mobile computer labs operated by stable teams of well prepared experts considerably simplified the implementation of testing, decreased its dependence on local conditions, eliminated potential negative local influences and increased its technical reliability and smooth execution.





www.oecd.org/edu/ahelo

engineering test session at the Faculty of Civil Engineering, Technical University Košice (April 24, 2013)

Key data on participation

Slovakia took part in all three strands of AHELO Feasibility Study.

Key data on the Generic Skills strand: 16 participating HEIs, out of which 11 were public HEIs, two state HEIs and three private HEIs. 1 541 students took part, representing 57.4 % of invited students. 378 faculties took part, representing 66.5 % of invited faculties.

Key data on the Economics strand: eight participating HEIs, out of which seven were public HEIs and one a private HEI. 929 students took part, representing 73.1 % of invited students. 132 faculties took part, representing 53.2 % of invited faculties.

Key data on the Engineering strand: engineering is taught at just three public HEIs in Slovakia, all of which participated. 358 students took part, representing 64.9 % of invited students. 90 faculties took part, representing 75% of invited faculties.

National and institutional management

Slovakia joined the AHELO project at the very last moment – in December 2010. The Ministry of Education, Science, Research and Sport of the Slovak Republic (the Ministry) perceived the project as a potential contribution to increasing the quality of Slovak higher education, which is the subject of long-term and very intensive debate, but which is based mostly on indirect information or on anecdotal evidence.

The implementation of the AHELO feasibility study in Slovakia (AHELO-SK) was defined by the Ministry as a development project. The Ministry decided to take part in all three strands. At the Ministry level, Peter Mederly was nominated as National Co-ordinator. After consultations with rectors and experts from universities that could potentially work as National Project Managers (NPM), the Ministry asked three HEIs to manage the single strands. After their approval, the Minister nominated Professor Roman Nedela from Matej Bell University in Banská Bystrica (UMB) to be the NPM for the Generic Skills strand, Associate Professor Ján Boďa from Comenius University in Bratislava (UK) to be the NPM for the Economics strand and Professor Ján Kalužný from the Slovak Technical University in Bratislava (STU) to be the NPM for the Engineering strand. Based on their requirements, the Ministry provided them with the necessary budget for the preparation and implementation of the project. The NPMs and their teams attended relevant training sessions organised by the OECD and the Consortium and prepared the translation of the tests and further materials during 2011.

At the beginning of 2012, the Ministry launched the Call for Participation in the development project AHELO-SK for Slovak HEIs. The Ministry promised the HEIs to cover the costs of the project. The subsidy for each HEI participating in a strand of the project consisted of a lump sum of EUR 5 000 and a further sum, the amount depending on the number of actual students and faculty taking part in the project (EUR 50 per student/faculty). To make the implementation as simple as possible for the HEIs and, at the same time, to make it reliable, the Ministry provided funding to the NPM for the purchase of mobile computer labs (53 notebooks for the Generic Skills strand and 33 notebooks for both the Economics and Engineering strands).

The reaction of the HEIs to the call was very positive and, in the end, 18 out of 33 Slovak HEIs took part in the project. One HEI – the Technical University in Košice – took part in all three strands, seven HEIs took part in two strands and ten HEIs took part in one strand.

The NPMs' teams managed their duties to a high standard. All NPMs had previous experience in the highest positions of university management (two of them had worked as vice rectors and one as a dean). They were able to hire very good Lead Scorers for their teams (Dr. Vladimír Poliach from UMB for Generic Skills and Associate Professor Peter Makýš from STU for Engineering – for Economics the NPM Ján Boďa also took the role of Lead Scorer) and other project managers. All NPMs had their team of specialists for the preparation and operation of the mobile computer labs so that no special technical training was necessary at individual HEIs.

Because of the volume of work (16 HEIs), the management of the Generic Skills strand was the most demanding. This is why the NPM's team for this strand had the most members. In addition to the NPM and Lead Scorer, there were further managers on the team: the head of the translation unit (Associate Professor Mária Spišáková) and the head of testing (Professor Štefan Porubský, current vice rector of UMB).

Of course, the NPMs' teams also used the infrastructure of their universities to complete their tasks.

Preparation for fieldwork

The preparation for fieldwork was organised according to the OECD and Consortium guidelines. Responsibility for this was undertaken by the heads of the relevant teams.

Fieldwork operations

There was a contact person at each HEI participating in the project. As a rule, it was a vice rector for education or another person from the HEI's top management. His or her responsibility comprised of assuring basic conditions for testing (time, room), and communication with students and faculties of the respective HEI. The sampling and training of team members were organised centrally. Academic information systems facilitated sampling, as well as communication. Using mobile computer labs eliminated the need for the training of local employees at HEIs.

Feedback from students/faculty

The students participating in the tests often expressed that they participated because they would like to compare their knowledge with students from other HEIs. At several HEIs, the rector or someone else from top management opened the test sessions in person. The participation was also considered by students as a kind of representation of the HEI. As regards the content, some students said that there was some difference between the emphasis in the AHELO tests and that in their study programmes. The first evaluation of the results indicates that, in some cases, there are relatively big differences between single HEIs.

Scoring process

Our experiences from the scoring process lead to several remarks and suggestions for the future. First, it could be useful to engage participating countries directly in the creation of the CRT. Next, the CRT should be a little simpler. Those used were too complex and this influenced the scoring process in a negative way. We would also recommend using simpler scoring rubrics for the CRT. At the scorer's training, more examples should be provided, not only typical ones. For the future, it would be very helpful to get the whole database of answers together with their assessment and also some data from the monitoring of the correspondence of scorers.

Results

The highest response rate was achieved in the Economics strand (average 73.1%, maximum 92.9%, minimum 32.5%, median 78.8%). The second highest was in the Engineering strand (average 64.9%, maximum 78.3%, minimum 50.5%, median 68.6%). For understandable reasons, the highest diversity and the size of the samples, the lowest response rate was in the Generic Skills strand (average 57.4%, maximum 96.3%, minimum 16.9%, median 55.3%).

As for the overall achievements in implementing the AHELO-SK project, we would like to mention the following:

- High participation of HEIs and high, in some cases even very high, participation of students in the testing.
- Obtaining new information on the Slovak HEIs based on methodology guaranteed by an international consortium.
- Successful organisational and technical management, including good decisions regarding the technical arrangements (mobile computer labs), leading to smooth implementation.

Impact at national/institutional/faculty level

We will definitely analyse AHELO data at different levels. But it is too early now for the formulation of concrete reflections.

Any particular innovative process you would like to share?

For a small country like Slovakia, the usage of mobile computer labs operated by stable teams of well prepared experts considerably simplified the implementation of testing, decreased its dependence on local conditions, eliminated potential negative local influences and increased its technical reliability and smooth execution.

Any particular challenge or problem you met?

As for the challenges, finding ways to motivate HEIs and students to participate in the AHELO project was one we were, at the beginning, perhaps most afraid of. Fortunately, it turned out that it was possible to motivate the HEIs and students to take part in innovative activities and to engage them with the offer of obtaining information on their results in order to compare them with those of their peers.

To take the risk of a relatively high financial investment in a project for which the development and results were uncertain at the moment of making the decision was the second challenge, related partly to the first one. We are happy that this challenge has also been met. The lesson for us from this challenge is that success in a project of this kind requires the synergy of several factors: the enthusiasm and professionalism of people; adequate funding; willingness to take risks and the support of the Centre.

While the first two challenges can be considered internal, the third also contains an external element. As we mentioned above, the Slovak Republic joined the AHELO feasibility study later than the majority of the rest of participating countries. The shorter amount of time was a big challenge for us. But it seems that it was a challenge also for the Consortium because some deadlines were not met on their side and it caused rather serious problems for us, mainly in the Generic Skills strand, where we had the highest number of participants.

Three suggestions for a main study

First, based on the results of AHELO feasibility study, we believe that, in principle, it is feasible to get external and internal comparisons of HEIs of different countries and therefore Slovakia supports the continuation of the AHELO process in the form of a main study.

Second, based on our experiences from the feasibility study, we recommend improving the preparation of tests and providing, in a feasible way, the possibility of comments by experts from participating countries. This recommendation comes from our experiences from training of scorers in Paris, where it turned out that the prepared tests contained questions where the responses were not clear even for the authors of questions.

Third, data acquired at the testing in a given country should be available to the country for further utilisation.

United States



Fieldwork, Analysis and Looking Ahead

With initial financial support from the Hewlett Foundation and subsequently from the U.S. Department of Education, three U.S. states and 11 American universities agreed to participate in the AHELO feasibility study Generic Skills strand. Beginning in 2010, the small, nonprofit association of State Higher Education Executive Officers (SHEEO), located in Boulder, Colorado, with two other non-profit higher education policy organizations, provided project co-ordination and international representation for U.S. participants, and sought to broaden understanding and support for the project among other American stakeholders. During 2012, working with the state higher education commissioners or system chancellors in Connecticut, Missouri, and Pennsylvania and institutional teams led by Institutional Coordinators (ICs), SHEEO organized and coordinated the population sampling, test administration, data reporting, scoring and other components of fieldwork by of the 11 American colleges and universities (several private as well as public) listed below.

AHELO Participants and Institutional Coordinators

- Central Methodist University, Missouri, Amy Dykens
- Cheyney University of Pennsylvania, Wesley Pugh
- Clarion University of Pennsylvania, Susan C. Turell
- Edinboro University of Pennsylvania, Rene Hearn
- Lock Haven University of Pennsylvania, David L. White
- Millersville University of Pennsylvania, Joseph R. Maxberry III
- Missouri State University, Kelly Cara
- Southern Connecticut State University, Michael Ben-Avie and Maureen Gilbride-Redman
- Truman State University, Missouri, Nancy Asher
- University of Central Missouri, Michael Grelle and Judi Reine
- Webster University, Missouri, Julie Weisman

Primarily regional, teaching –focused institutions, these 11 American universities can make no claim whatsoever to be representative of the more than 4 500 diverse colleges and universities across all 50 states. U.S. participants were also a small part of the 97 institutions across nine nations participating in the Generic Skills strand. It should be noted that the states and institutions volunteered to participate out of their own self-interest, and that they received only minimal financial reimbursement to cover approved, out-of-pocket project expenditures while providing significant in-kind support. Using the international manuals and guidelines, participating U.S. institutions assessed a sample of their graduating students using an instrument consisting of one of two rotating performance tasks adapted from the Collegiate

Learning Assessment [CLA] and a set of complex selected response items provided by the Australian Council for Educational Research (ACER).

Several of the U.S. institutions had prior experience administering the CLA and all had some prior exposure to broad-scale learning assessment; because of this experience and exposure, the test administration components of fieldwork posed few technical or administrative challenges. All institutions also had centralized student, faculty and institutional data bases, and professional staff capable of preparing the sample population files. The contextual, faculty and institutional data collection and reporting also posed few if any challenges for U.S. participants. The primary challenges of the fieldwork resulted instead from the extremely compressed institutional preparation and implementation timelines, the limited content and specificity of the international field manuals, and the complexity involved in accessing the numerous secure international websites.

From late 2011 and through the first five months of 2012, short and tight international timelines for the fieldwork placed considerable pressure on all U.S. participants. Final versions of international fieldwork guidelines and manuals were not available until mid-December 2011. Although Institutional Co-ordinators for the fieldwork were designated earlier, fieldwork “teams” could not be organized until the essential steps, components and scheduling of the fieldwork were made available. In addition to the IC (typically designated by the Provost), the institutional teams generally involved individuals from Institutional Research, communications, test administration or computer lab facilitators, interested graduate students, and high-level representation and support from the President or Provost.

Charles Lenth, the designated National Project Manager (NPM) from SHEEO, met with the state and institutional co-ordinators in Missouri in mid-December 2011, with the entire institutional team and state co-ordinator in Connecticut in earlier January 2012, and with two groups of Pennsylvania institutional co-ordinators later in January-February 2012. Each meeting included an introduction to the feasibility study, review of the international fieldwork manuals and guidelines, and discussion of student and faculty sampling, test administration, timelines and other aspects of the fieldwork.

The institutions prepared and provided to SHEEO sampling frames for both students and faculty eligible to participate in the Feasibility Study (essentially descriptive data on the “in-scope” populations). SHEEO contracted with the National Center for Higher Education Management Systems (NCHEMS, with whom SHEEO shares facilities in Boulder, CO) for technical support in reviewing the sample frames, resolving any questions with the institutions, and drawing the random samples using the software and routines prescribed by international guidelines. Both the sample frames and the randomized sample files were then sent to Statistics Canada (the ACER Consortium member responsible for sample design and weighting) for review and quality control. When approved, the data files were returned to SHEEO with assessment access codes added for all students in the sample and “electronic tokens” for faculty in the faculty sample. After final inspection, SHEEO returned these data files to the institutions for their use in test administration and data collection.

Simultaneously with the sampling steps, institutions began to schedule student assessment sessions, prepare for test administration, and undertake other fieldwork components. Student assessments began at the participating U.S. institutions in late February and continued through early May. Some institutions scheduled 2-3 larger testing sessions and others numerous smaller sessions, depending primarily on the capacity and availability of computer labs or testing facilities. All sessions were proctored according to international guidelines, with students provided individual codes to allow direct but controlled access to international testing websites. All responses were recorded directly on these sites. To resolve any questions that arose during testing sessions, SHEEO provided assistance to the institutions as necessary and received technical support through the 24/7 emergency help desk maintained by ACER and accessible by the NPM.

The faculty survey was administered during this same time period, usually initiated by a letter of invitation and encouragement from the president or provost emphasizing the importance of institutional and faculty participation in AHELO. These letters contained the “tokens” so that faculty could complete the AHELO questionnaire at their convenience by logging into a secure international website.

All participating U.S. institutions took steps to inform students and faculty about AHELO and to encourage participation. These communication steps included institutional or system press releases, email communications, posters, and other promotional approaches. Sampled students were encouraged to participate through letters of invitation sent by institutional leaders or by state commissioners or system chancellors. All institutions also provided some form of incentives to students who participated in the assessment; these incentives varied from waiving institutional graduation fees or preferential family seating for graduation ceremonies, to cash or gift card rewards (generally about USD 50, but at one institution up to a maximum of USD 100). No monetary incentives were provided to faculty, although most institutions provided repeated and personalized encouragement to their sampled faculty. All promotional steps and participation incentives were determined by the institutions and then reported as part of an institutional context questionnaire. SHEEO also requested all institutions to maintain an AHELO participation logbook to record all fieldwork as it was completed.

Scoring of the student assessment performance task for U.S. institutions was done under the direction of an experience CLA scorer recruited by SHEEO, and under contract with SHEEO participated in the face-to-face international scorer training provided by ACER in March. The Lead Scorer recruited and trained five additional Scorers for scoring that commenced the third week of May. All student assessments were double scored in a randomized manner and major discrepancies in these scores were identified and resolved by the Lead Scorer. The Lead Scorer also monitored all scoring using procedures adapted from CLA scoring. International consistency in scoring was checked by translating student responses into different languages and feeding a sample of translated responses into the scoring queues of selected nations, including the U.S. The international windows for test administration and data submission ended on 30 May and all U.S. scoring was completed by 20 June.

Tallies and frequency tables on student and faculty participation in AHELO were returned to SHEEO and then to the institutions in August. In total, 734 students participated in the student

learning assessment, from a total sample population of 2 296, with all but 17 of these providing complete and usable data. The median student participation in the Generic Skills assessment across all 11 U.S. institutions was approximately 30%, with a range of 67% to less than 10%. This compares to a nine-nation median Generic Skills participation rate of nearly 53%, with a range of 95% (Columbia) followed by 78% (Mexico) to a low of 12% (Finland) and 8% (Norway). A total of 324 faculty participated in the faculty survey out of a total sample of 523, with 317 of these deemed usable. The response or completion rate for faculty overall was just over 60%.

Among the U.S. participants, both student participation rates and the degree of challenge institutions faced in completing the fieldwork appear to be influenced by prior experience with large-scale learning assessment. Four of five Missouri institutions, where there has been a “culture of learning assessment” for nearly two decades, achieved participation rates higher than all five Pennsylvania institutions, where experience with learning assessment was generally more recent. U.S. faculty participation rates were consistently higher than student participation across all 11 institutions, with a range of 45-83%.

Prior to and following the final meetings of the AHELO GNE and Stakeholders Group in October, SHEEO participated in the review and revision of multiple drafts of the feasibility study findings and reports. In late December, national and institutional student assessment data files were returned to SHEEO. These data files are currently being prepared for return to individual institutions along with a standard international participation report. SHEEO is planning to work with the participating states and institutions, and use the expertise of NCHEMS, to analyze U.S. assessment data within OECD’s guidelines for data use, interpretation and public access.

SHEEO is also keen to explore the potential to share national data and data analyses with other interested, participating nations. The U.S. delegation to the feasibility study Conference and Symposium will include the State Project Co-ordinators from Connecticut, Missouri and Pennsylvania along with representatives of the U.S. Department of Education, the U.S. Delegation to OECD, and SHEEO, where we hope that opportunities for such data sharing can be discussed and agreed upon.

If the findings and results of the Feasibility Study convince the OECD Education Policy Committee to proceed to a larger scale AHELO programme, U.S. participants believe that several important organizational and managerial observations will need to be taken into account. In brief and based on our involvement in the feasibility study, these observations include but are not limited to the following:

5. Any future, large scale AHELO-type assessment must clearly, convincingly and realistically specify the purposes and limitations of cross-national learning assessment at the level of college graduates in order to avoid the misunderstandings and unfulfilled promises observed or experienced by feasibility study participants.
6. International project leaders and managers must determine in advance that financial resources and support will be adequate; that project budgets are appropriately structured, managed, and monitored, and that national commitments are secure from the outset of the project in order to avoid, to the extent possible, the delays and changes encountered during the feasibility study.

7. There must be a thorough and reliable international pilot test of all assessment instruments and test administration systems, followed by any necessary modifications to ensure that they are a good fit for the highly variable national and institutional contexts in which they will be used.
8. A broad-scale international assessment will need to take full advantage of international contractors, researchers, and assessment experts in both designing and implementing the overall project, while avoiding the inter-agency conflicts and competition that hindered the feasibility study and making sure that the best available expertise is used well and wisely.
9. Such a project will need strong leadership and careful management by OECD or another appropriate international agency, appropriately placed and professionally supported within that organisation.

While U.S. participants in AHELO have learned from and, we believe, contributed to the Feasibility Study, our willingness participate in an AHELO-type programme moving forward will require a careful and open discussion of the lessons learned at both the national and international levels in order to achieve reasonable agreement on the purposes and value of such an investment of time, money and limited educational resources over an extended period.

NOTES

1. CAE Performance Academy Proposal may 2010.
2. This goes for the small-scale validation of the assessment as well as the assessment in phase 2 of the Feasibility Study.

CHAPTER 9

ROLE OF THE AHELO FEASIBILITY STUDY TECHNICAL ADVISORY GROUP (TAG)

Peter T. Ewell

The purpose of this chapter is to describe the creation, organisation, and role of the Technical Advisory Group (TAG) during the AHELO feasibility study. The first section describes the creation of the TAG and how its role evolved as the feasibility study developed. This section also examines TAG operations — how meetings were conducted and the nature of the Group’s ongoing work. An extensive second section provides a thematic analysis of the principal issues the TAG considered and the most important recommendations it made to help guide the conduct of the feasibility study. A third section presents the TAG’s recommendations about the conduct of any future AHELO Main Study. The fourth, and final, section provides the TAG’s assessment of the conduct of the feasibility study as a whole. The TAG is also charged with providing the GNE with a definitive recommendation on the feasibility of an AHELO Main Study, but cannot do so until all relevant evidence has been presented and thoroughly reviewed. Because this is not yet the case, the chapter does not provide such a recommendation.

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 recognised by the OECD Secretariat and interested countries from the outset of the AHELO initiative. Similar bodies had been created for both PISA and PIAAC—the OECD initiatives most comparable to AHELO — and had proven useful.¹ This need was affirmed by the three expert group meetings held in 2007 in Washington, Paris, and Seoul. The membership and conduct of these expert group meetings, moreover, in many ways resembled the eventual role of the TAG although their principal task was to establish the need for and plan AHELO. Many of these experts were assessment specialists and the topics they considered — appropriate unit of analysis, the subject domains for assessment, the timing of the assessments, the broad properties of the assessment instruments to be used, and how results should be analyzed and reported — were the kinds of topics subsequently addressed by the TAG.

The TAG was formally established in 2010 with eight members drawn from assessment and higher education experts throughout the world.² In the subsequent three years, two members of the TAG resigned because of other pressing commitments and were replaced.³ The TAG reported to both the Secretariat and the AHELO GNE and was initially managed by the Consortium in its capacity as the overall manager of the feasibility study. The Terms of Reference of the TAG (Annex D) 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 any 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 Contextual Dimension. This meant a particular focus on the technical challenges of conceptualising generic skills and closely reviewing instrumentation and analysis in this domain area, as well as comprehensively reviewing the construction and implementation of the three context surveys. 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. In order to further establish the TAG's independence in this role, management of the TAG was shifted from the Consortium to the OECD Secretariat in early 2012.

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.⁴ Agendas for each meeting were developed in partnership with the Chair of the TAG by the Consortium or the OECD Secretariat, depending on which of the latter was responsible for management of the TAG. 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 co-ordination, assessment administration procedures, scoring arrangements for constructed response tasks (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. Ongoing contact between the two Chairs proved essential in surfacing and resolving implementation issues at several important junctures in the course of the study.

From the outset of the feasibility study, the international financial situation and its impact on the study's budget affected the activities of the TAG. As noted, only three of the eight TAG meetings could be held in a face-to-face setting. This meant significant limitations on what could meaningfully be undertaken and accomplished. It also meant that the TAG Chair had to devote more time than anticipated to managing issues that could not be properly attended to in meetings held by teleconference, where frequently not all TAG members could attend simultaneously because of global time zone differences. Budgetary limitations also constrained the ability of the Consortium and the OECD Secretariat to follow through on many TAG recommendations. Indeed, comments about the growing danger of budget constraints threatening the ability of the feasibility study to reach valid conclusions are a prominent theme across TAG reports.

Major TAG recommendations made during the Study

The TAG provided scores of substantive recommendations to the Consortium and the OECD Secretariat over the course of the feasibility study. An analytical reading of reports from the eight TAG meetings to date, however, reveals that many of these recommendations fall under a few broad themes, which were repeatedly raised and discussed by the TAG. This section therefore reviews major TAG recommendations under a range of thematic headings. These themes not only serve to cluster the TAG's recommendations in a way that is easily understandable, but they also signal the major concerns the TAG had (and continues to have) throughout the study.⁵

The importance of contextual information

The TAG emphasised the importance of contextual information (Module D) in its initial meeting in September 2010 and continued to do so in virtually every subsequent meeting. The TAG believed that the collection of contextual information was imperative for several reasons. First, in the absence of contextual information about differences across instructional settings and student experiences, members feared that AHELO would become little more than a ranking of un-interpretable numbers. Without such information, benchmarking and comparison across institutions and programmes — the principal purpose of any AHELO Main Study — cannot be undertaken meaningfully. Taking such variations in context into account in any analyses of the results of the three cognitive assessments is especially important in an international setting where such differences involve significant variation. Second, the contextual data collected through institutional, faculty, and student surveys was the only set of data collected for all strands across all institutions participating in the feasibility study. Although the TAG recommended on several occasions that a common core of cognitive items be included in all three assessment strands, the student context survey supplied the only set of data that could be used to link the three assessment strands.

As a result, the TAG repeatedly expressed concern that the Contextual Dimension might be scaled back to save resources in the light of steadily deteriorating project budget conditions. These concerns were exacerbated by early scepticism expressed by some members of the GNE about the validity of answers about instructional contexts supplied by students through a survey and about the overall utility of such information in guiding improvement in relation to the burden of collecting it. That said, the TAG made a number of recommendations intended to reduce this burden. First, it strongly resisted temptations to add items to the contextual surveys that would be interesting to know, but that would not be immediately useful in interpreting results of the cognitive assessments. Second, it urged the Consortium to make use of existing information from national sources and from such international initiatives as U-rank and U-Multirank that were readily available. Third, it urged the Consortium to prioritise contextual data collection to emphasise the surveys that would provide the most important contextual information. The Student Survey administered in conjunction with the three assessments was deemed essential in this regard, with the Institutional Survey considered important, and the Faculty Survey of lesser importance. Fortunately, despite funding shortfalls, all three surveys were administered as planned.

As the feasibility study evolved into Phase II, moreover, the TAG became increasingly interested in pointing out less tangible variations in context. By this time, all of the Module D surveys had been finalised and data collection was about to proceed. But discussions of project operations including reports on contextualising the assessment instruments, training scorers and those responsible for administering the assessments, and early cognitive interviews and try-outs were beginning to reveal more subtle variations in context not captured by the formal survey instruments. For example, the process of building the assessment framework in Economics with the assistance of an international expert group in the discipline revealed significant and previously unknown differences in the way the discipline was conceived and taught across different countries. In its February 2012 meeting in Tokyo, therefore, the TAG expressed its

belief that deliberate efforts needed to be undertaken by project leadership and country teams to capture such important pieces of “tacit knowledge” about individual country contexts and participating students that were not formally documented by the Module D Surveys.

Scope and Budget

The AHELO feasibility study was an ambitious project but, as originally conceived through the three Expert Group Meetings held in 2007, it was intended to be limited in scope. The conclusion of the Seoul meeting, for example, was a consensus that the feasibility study should include “at least three countries in three languages,” and that assessments be administered at “from three to five institutions in each country (OECD, 2012, p.70).” By the time the study was underway in 2010, however the global fiscal crisis had severely limited the budget for the initiative. This provoked concerns from the TAG from the outset about whether the study could provide valid and usable results with limited resources. But it also compelled the OECD Secretariat to expand country participation, in part in order to collect more participation fees to move forward with the study. Several participating countries joined the study quite late, and were consequently relatively underprepared, compared to early country participants. Eventually, of course, a total of seventeen countries and 248 institutions took part. By its first face-to-face meeting in Paris, which took place in April 2011, the TAG was seriously concerned about the growing scope of the study and recommended a) that no further countries be added and b) that the study be scaled back to three or four countries per assessment strand and ten institutions per country. The former recommendation was eventually followed and a firm cut-off date was established for country and institutional participation in early 2012. The latter recommendation was not followed, but the OECD Secretariat reports that some important insights were gained through the participation of late-joining countries so, on balance, the decision to add countries added value.

Because the scope of the feasibility study was rapidly approximating that of a full-scale international assessment effort, the TAG also expressed early concerns about managing expectations about what it could accomplish. As early as its initial meeting in 2010, the TAG commented on the Consortium’s *Assessment Design* document (AHELO Consortium, 2010) that cautions for stakeholders should be inserted stressing the fact that the feasibility study would not produce valid country-level results. The TAG also noted that the document should more clearly communicate that although usable institution-level results might be produced, the primary intent of the feasibility study was to determine if it was possible to **implement** such an assessment effort at the tertiary level across diverse country and institutional contexts to yield valid results. In a related vein, the TAG went on record in its April 2011 meeting that it was important to prevent the governments of participating countries from expropriating assessment results for use in high-stakes reward or punishment schemes like performance funding. Finally, the TAG stressed that while it was appropriate not to disclose results in the early stages of the implementation of AHELO because of uncertainties about the validity and reliability of the assessments, maximum transparency about the conduct of the study should characterise its implementation and that results should be disclosed as soon as feasibility had been established.

Similarly, the TAG believed that all of the lessons learned in conducting the feasibility study should be quickly and clearly disseminated at the conclusion of the project and that these lessons should eventually be integrated with parallel international initiatives directed at improving instructional quality such as qualifications frameworks, subject benchmarks, Tuning, and U-Map and U-MultiRank. As scholars of the field, members of the TAG saw the results of all such projects coming together to constitute a “new policy narrative” on learning outcomes that might transcend purely descriptive benchmarking and assessment efforts. Consequently, its members urged stakeholder and higher education policy groups to come together to shape this wider narrative at the conclusion of the study. The AHELO feasibility study conference in March 2013 will provide one opportunity to accomplish this important goal.

As a final budgetary note, the TAG called attention several times to the balance of funding allocated across modules for instrument development across modules. As of the April 2011 meeting in Paris, development and implementation allocations for Module A, which involved two pre-existing open response instruments slightly exceeded those of Modules B, C, and D combined, which involved four purpose-built assessment instruments and three contextual surveys, as well as all the study management and co-ordination activities included under Module E. This imbalance was the result of the original procurement and tendering process that preceded creation of the TAG and was later largely rectified as the study’s budgetary challenges were addressed.

Sampling and response rates

Like any assessment, the validity of AHELO depends heavily on the provision by institutions of a representative sample and their ability to obtain a usable response rate in the form of students who actually took the assessments. Indeed, one goal of the feasibility study itself was to determine how well institutions in differing educational settings and cultural contexts could fulfill these tasks. As a result, the TAG reviewed numerous documents on sampling and commented on this topic on several occasions.

In its second meeting in December 2010, the TAG discussed the need for strong student sampling methodologies, which were nevertheless flexible enough to retain institutional engagement. But the consensus of TAG opinion was that implementing scientific sampling should take precedence over potential objections from participating institutions who wanted to draw their own samples or who wanted to participate in the study but could not supply an adequate sampling frame. As it happened, more than a few participating institutions ended up proceeding with the study using convenience samples from which obtained results could not be validly generalised. This led to a recommendation by the TAG that a relative aptitude measure be included in the sampling frame wherever possible, to ensure that institutions were not selecting their best students for testing and that aptitude information should be reported for non-respondents. The TAG also recommended that the Consortium closely monitor institutional use of stratification to ensure that properly representative samples of students were chosen.

Turning to response rates, the TAG was equally concerned that these be high enough to ensure adequate generalisability, but believed that the 75% requirement set by the Consortium was

not realistic for many country contexts. While TAG members endorsed the objective of achieving as high a response rate as possible, they believed from experience in surveying or testing university students that this level of response would be unlikely. As it happened, in this case members of the TAG — drawn largely from Europe, America, and Australia — were constrained by their own cultural contexts, because several countries actually exceeded this response rate target by wide margins. Nevertheless, response rates varied substantially across national contexts and sometimes across institutions within a single country context. As a result, the TAG eventually endorsed broadly the Consortium's decision to set a minimum response rate standard of 50% for including data in analyses, with the caveat that in the future, empirical evidence should be sought to justify such a threshold in terms of the reduction of non-response bias. The TAG also pointed out that actual response rates obtained varied because of institutional commitment to the project and how the testing and survey instruments were organized; more detailed attention to these issues might therefore yield better response rates in any AHELO Main Study.

A related issue discussed by the TAG at several points in the study was the use of incentives to motivate students to participate in the assessment and do their best. Here the TAG's concern was that the use of different kinds of incentives (especially if they were differentially effective) could distort the obtained response, rendering the obtained results less comparable. As a result, the TAG recommended that an item should be included in the Module D Institutional Survey to ask ICs whether incentives were used and, if the answer was affirmative, what particular incentives were used. At the same time, the TAG suggested changes to the list of allowable incentives included in the *Test Administration Manual* prepared by the Consortium (AHELO Consortium, 2011).

Consistency of implementation across countries

Another potential threat to the validity of international assessments is inconsistent implementation across countries. Examples of such variations include different test administration procedures, variations in sampling and incentive schemes (as discussed above), and inconsistencies in the scoring of constructed response tasks (CRTs) that all three assessments contained. Throughout its review of the documents produced by the Consortium to guide implementation including sampling materials, test administration manuals, scoring instructions, and score reporting procedures, the TAG was particularly concerned about consistency of language and cross-referencing between documents. In the course of several of its meetings in 2010-2011 it pointed out potential inconsistencies to the Consortium which were rectified.

Probably the most important area of concern for the TAG under this heading was ensuring consistency in the manual scoring of CRTs, especially in the Generic Skills strand. In its September 2011 meeting, for example, the TAG agreed that rigorous scorer training was the key to ensuring comparability across countries but recommended that vignette analysis be explored and that a variety of different "anchor papers" be prepared for each level of the scoring rubrics to illustrate the intended level of performance. By its face-to-face meeting in Tokyo in February of 2012, some of these concerns became more pointed after an overall favourable review of detailed reports of the various translation, contextualisation, and

adaptation procedures used for Modules A-D. Specifically, the TAG was concerned that the procedures used for scoring CRTs in all three assessments were excessively local. Each country would be undertaking its own scoring under the direction of its Lead Scorer in isolation and this might subsequently limit the generalisability of a given country's results beyond its own borders. To help address this, the TAG recommended that the Consortium establish a scorer query system that would enable cross-country variation to be moderated and documented, as well as convening scorers from multiple country contexts in virtual focus groups after scoring has occurred. Because of budget and logistical constraints, this was not done.

A particular concern of the TAG was the effect of translations of student responses in the CRTs in Generic Skills on the generalisability of the resulting scores. Accordingly, in its February 2012 meeting, the TAG recommended that the Council for Aid to Education (CAE) — the contractor responsible for developing and administering the Generic Skills CRTs — proceed with two proposed studies intended to examine the effect of translation on score results. The first of these studies involved translating into English a sample of student responses from the language in which students answered them, then re-grading these English-language answers. The second involved the reverse: translating a set of English-language answers into the language of another country participant, then asking graders in that country to score the responses. Both studies were accomplished and helped bolster confidence in the validity of the translation process used in the Generic Skills CRTs. An additional special study recommended by the TAG was intended to determine whether or not students' disciplinary background had an impact on Generic Skills CRT performance. Previous studies directed at this question undertaken by CAE for the Collegiate Learning Assessment (CLA) in the U.S. — the instrument on which the AHELO Generic Skills CRT was based — indicated that no such discipline effect was present, but the TAG wanted to be assured that this was also the case in an international context. For a variety of reasons, and much to the disappointment of the TAG, this study was not done.

Finally, the TAG noted favourably the Consortium's decision to retain an independent organisation — the Data Processing and Research Centre of the International Association for the Evaluation of Educational Achievement (IEA-DPC) — to conduct evaluations and audits of Phase II data collection activities. The presence of this external organisation, in the TAG's view, did much to assure confidence in the consistency of study implementation across diverse country contexts.

Generic Skills

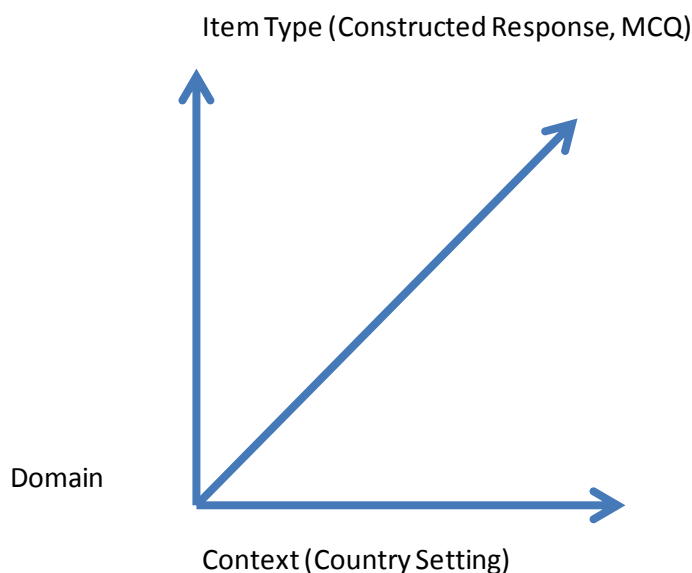
Another important theme of TAG discussion over the course of the feasibility study was the wisdom of assessing Generic Skills as an independent endeavour. The TAG noted that many assessment professionals dispute the validity of assessing such competencies as critical thinking and communication in the absence of a disciplinary context — a view shared by some country participants and members of the GNE. The TAG noted that both the Engineering and the Economics assessments contained some "generic" items in their CRTs, and this might provide an opportunity to investigate this matter more directly. In the same vein, the TAG noted that some countries were administering all three assessments and, as also discussed by the GNE and the Stakeholders Consultative Group (SCG), this might provide some opportunity for cross-testing: administering either the Economics or the Engineering assessment along with the

Generic Skills assessment to the same body of students. For budgetary and logistical reasons, this idea was not pursued.

The TAG also made important recommendations on the structure of the Generic Skills assessment itself. The original design for this assessment was similar to that of the two disciplinary assessments — a single CRT based on the CLA and a specially-developed battery of multiple-choice questions (MCQs). Midway through the study, the development of the MCQ component was dropped. The TAG expressed serious concern about this matter in its April 2011 meeting, noting that the MCQs would have enabled the equating of CRT results across different contexts and countries. The TAG also noted that the results from the MCQs would have provided a valuable check on the rubric-based scoring of CRTs. This check would be particularly important in light of the fact that all CRT scoring would be accomplished locally for each country. While the TAG's recommendation was not in itself decisive in driving this decision, an MCQ battery supplied by the Consortium was reinstated in the Generic Skills assessment, albeit one not developed explicitly for AHELO.

Analysis of results

With testing immanent in the spring of 2012, the TAG made a number of recommendations about how the resulting data should be analyzed. In its February 2012 meeting, the TAG proposed that analyses be undertaken of the parallel performance of constructed response prompts and MCQs across all three domain areas (Economics, Engineering, and Generic Skills), and across different contextual conditions. As portrayed in the accompanying diagram, the TAG believed that this three-way design should constitute the basic analytical approach for interpreting results drawn from throughout the feasibility study. As such, it could very well serve as the overall framework for organising the final report on the feasibility study. Wherever possible, moreover, the TAG recommended that all such analyses should be undertaken at the sub-score level. Finally, the TAG observed that the posture of these analyses should be cast as a set of “quasi-experiments” to examine the question of what combination of the three variables yield optimum results. The TAG also recognised that data limitations might mean that some interactions might not be feasible.



When the TAG reviewed the drafts of the Consortium's final report in the advance of its October 2012 meeting, it again urged that this analysis design be used and that all implied analyses be undertaken and reported. This recommendation was reinforced by a memo from one member of the TAG to the Consortium in the wake of the October meeting. As it happened, the majority of recommended analyses were included in the final draft of the Consortium's final report supplied in December, though not presented in the recommended format. The TAG is pleased that the OECD Secretariat has chosen to report results in this format, insofar as results are available, in the current volume.

Post-implementation activities and reporting

Finally, the TAG made a number of recommendations about what should happen to conclude the feasibility study. First, the TAG recommended at its February 2012 meeting that the Consortium should design and implement a post-implementation survey for NPMs and ICs to be administered after all data is collected. The principal purpose of the survey would be to determine after the fact and document any variations from sampling and test administration procedures that may have occurred in the course of implementation. At the same time, the TAG recommended that the expert groups for the Economics, Engineering, and Contextual instruments be reconvened retrospectively to consider any problems encountered or identified during the process of instrument contextualisation and administration. Differences in perceptions or beliefs among participating experts, the TAG believed, would provide important evidence to inform final evaluations of validity and reliability. As it happened, the Consortium asked IEA-DPC to survey NPMs and ICs and the results of this survey were helpful in identifying matters that would need attention in any AHELO Main Study. The second recommendation was discussed with interest by several of the Consortium contractors but has so far not happened.

After reviewing draft chapters of the OECD Final Report on the feasibility study in December 2012, the TAG also made a number of recommendations about how the OECD Secretariat should handle the conclusion of the study. First, it recommended that the Final Report be issued in stages, with a first volume reporting on the implementation of the feasibility study and subsequent volumes presenting results and conclusions. Second, it concurred with the Secretariat's intention of re-positioning of the conference planned for March 2013 from an emphasis on dissemination to an emphasis on lessons learned. Both of these occurred as anticipated.

Recommendations on the conduct of an AHELO Main Study

Despite the fact that no definitive conclusions about feasibility can be made at this time, the TAG can make some recommendations on possible features of a Main Study based on the data reported so far and the many lessons learned about implementation. This section provides the TAG's conclusions and recommendations about a possible AHELO Main Study. The first section addresses what the TAG considers to be the most critical commitments and design choices that the OECD must make in moving forward. The second section offers a series of important, but somewhat less crucial, recommendations about any future implementation of AHELO.

Critical commitments and design choices

A first set of TAG recommendations concerning the design of an anticipated AHELO Main Study can best be framed in terms of a set of critical commitments and design choices around four topics, as noted in the October 2012 TAG report.

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.

Costs

Two primary questions will probably be raised by any country/system considering whether or not to join a future data collection effort like AHELO: "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. Some of this information was collected through the surveys

conducted by IEA DPC, but some remains undocumented. Finally, no effort has as yet been made to ascertain the perceived opportunity costs associated with participating in AHELO. As a consequence, the TAG recommends that a systematic effort to collect data about both direct and indirect costs be included in any future Main Study. Not only will potential participants be in a better position to make a choice about whether to join the study but the OECD will have obtained important information on the return on investment associated with such an enterprise.

Project Management

Any future AHELO Main Study will require a management structure and set of enforceable contracts among data providers and participating countries/institutions that makes mutual transparency and active cooperation a condition. It will also require senior management to exert proactive ownership of the enterprise. Within the OECD context, this means that any future AHELO needs to be at least as actively managed by top level positions as is currently the case for PISA or PIAAC.

Instrumentation

The AHELO feasibility study provided an unprecedented opportunity to try out a range of approaches to collecting valid data on student learning outcomes in an international context at the higher education level. The three domain strands — Economics, Engineering, and Generic Skills — together with the all-important Contextual Dimension, represent a broad range of content and were actualised by means of a varied array of assessment designs. The information generated by the study thus provides a good basis for beginning to draw conclusions about appropriate instrumentation for future large-scale international assessment work in higher education. Based on results so far, the TAG believes that there are two major choices with regard to instrumentation to be made in moving forward.

The first choice is whether or not to include a dedicated Generic Skills strand in any Main Study. As the TAG indicated in several of its reports, the existence of these competencies independent of discipline or field of study is a contested issue in the field of higher education assessment. While the TAG believes that 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 contextualised. Of course, the TAG recognises that the two Generic Skills CRTs used in the feasibility study were contextualised to a “real world” problem-solving situation. Unfortunately, the results on generalisability reported in the Consortium’s report appear to indicate that the manner in which these tasks were culturally situated and perceived varied substantially across countries and systems.

How appropriate contextualisation 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 component, moreover, the TAG believes that the performance tasks constructed should be situated in the context of broad disciplinary groupings like the sciences, social sciences, humanities and fine arts. This was the architecture of the ETS *Tasks for Critical Thinking* from which the CLA — the basis for the Generic Skills CRTs used in the AHELO feasibility study — was originally derived.

The second choice about instrumentation is whether production based CRTs should be included in any future AHELO Main Study at all. Decades of research suggest that CRTs will never perform as well in terms of reliability as a battery based solely on MCQs. Results of the Consortium's report confirm this conclusion for all three domain strands, which is not surprising. The question for an AHELO Main Study is whether the use of CRTs adds enough validity to pay this inevitable price in lost reliability.

The TAG does not yet have enough information to provide a definitive recommendation about this choice. But results of the feasibility study in Engineering and Generic Skills suggest that something valuable was learned from the inclusion of CRTs. Indeed, discussions during the October 2012 GNE Meeting in Paris indicated that some of the most important information that could drive improvement in teaching and learning was obtained through the CRTs. Finally, the TAG recognises that the AHELO feasibility study provided only a limited opportunity to test the efficacy of CRTs *per se*. The Generic Skills CRTs were based on a particular example of a CRT — the CLA — that has unique features not shared by other possible CRTs. It can also be argued that the Economics and Engineering CRTs did not undergo an adequate test because they proved so difficult that a substantial proportion of students could not complete them.

The major drawback of including CRTs is substantially increased costs. On balance, the TAG supports the continuing use of CRTs in AHELO. In doing so, however, it reminds stakeholders that the technical construction of any assessment depends on its purpose. 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 a choice based solely on MCQs may be preferred.

Finally, the TAG strongly recommends that a common core of ten to twelve MCQs focusing on generic skills be included in every assessment administered as a part of AHELO. Doing so would provide an invaluable tool for equating, or otherwise comparing, other assessment results across institutions and system/country contexts at minimum additional cost.

Additional Recommendations

In addition to the major commitments and design choices reported above, the TAG can at this point definitively recommend a number of additional features that should be included in any AHELO Main Study. Many of these affirm recommendations advanced by the *AHELO Feasibility Study Report* prepared by the Consortium. They include:

Assessment Frameworks and Expert Groups

All assessments included in AHELO should be constructed on the basis of an accepted Assessment Framework. Each Assessment Framework should reflect broad international consensus about the scope and content of the domain to be assessed as reflected in published learning goals inventories, qualifications frameworks, and accreditation/licensure standards, as well as expert opinion. In the AHELO feasibility study, this was the case for Economics and Engineering strands, as well as the Contextual Dimension, but not for Generic Skills. Similarly, all assessments should be constructed under the direction of an independent Experts Group. The role of this Group begins with establishing and validating the Assessment Framework for the discipline or area. It should also include reviewing successive drafts of the instrument itself (as was done in the feasibility study) and reviewing results of the assessment *after* it has been administered (which was largely not done in the feasibility study). Finally, sufficient resources should be allocated to support the necessary number of face-to-face meetings of all such groups, including the TAG.

Faculty Survey

The TAG takes note of the Consortium's conclusion that the Faculty Survey component of the Contextual Dimension not be used in any future AHELO Main Study because of the inability to link faculty responses to student responses, the low response rates experienced, and the fact that its results were not broadly used in analyses. The TAG also recognises the fact that every additional survey involves added costs. But the TAG has reservations about the lack of faculty testimony in a study whose major purpose is the improvement of teaching and learning. It therefore recommends that this issue be further studied with an eye toward building a new faculty survey for AHELO focused primarily on faculty perceptions of the institution's teaching and learning environment. If the faculty survey is kept, moreover, its content should be better aligned with that of the student survey.

Readiness Criteria

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* (AHELO Consortium, 2012). The TAG believes that the *Technical Standards* themselves, periodically updated on the basis of experience, should be a permanent part of AHELO.

Field Trial

A full-scale field trial should be conducted prior to actual testing. The AHELO feasibility study could only include limited numbers of focus groups and cognitive interviews involving a very small number of students. In addition, a pre-scoring calibration study should be undertaken to identify countries/systems that are not scoring reliably so that they can be given additional training.

Quality Monitors

A project-wide Quality Monitor should be established, as well as a National Quality Monitor for each participating country/system. This is consistent with international standards in conducting such studies. The TAG believes an effective quality monitoring system could have prevented or mitigated some of the violations of study protocols that occurred in the AHELO feasibility study such as the lack of a population/sampling frame, non-random sample selection, inadequate training of in-country scorers and improper test administration procedures. The TAG also commends the Consortium's decision to retain an independent third-party contractor to carry out a quality control assessment and believes this should be a feature of all future AHELO assessment activities.

Sampling

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 clear to the TAG that there were any insuperable obstacles to constructing a sampling frame of eligible students, nor to drawing a probability sample. For any Main Study, participating institutions should be required to compile a list (or lists) of eligible students (or groups of students) and that either probability sampling or a census be employed. The TAG also believes that 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.

Response Rate Standard

The Consortium recommends that analyses only be performed using data from census or random sampled institutions that achieved at least a 50% response rate. The TAG believes that it may be reasonable for AHELO to impose a fixed minimum response rate threshold, at the level of the country or the institution, for inclusion in the analysis. However, the TAG also recommends that empirical evidence be sought to justify such a threshold in terms of the reduction of non-response bias. For example, it seems quite possible for each institution to assemble some data on student achievement (for example, grades or test scores) for both responding and non-responding students that could enable such a determination. For the most part, response rates were acceptable, but this does not mean they cannot be improved. Accordingly, the TAG also believes that measures to increase response rates should be actively researched before any new AHELO data collection.

Post-implementation survey

The OECD Secretariat should conduct or contract for a post-implementation follow-up survey of NPMs and ICs participating in any future AHELO Main Study. This survey should include questions on the perceived value of AHELO and especially on the use of assessment results to discuss and implement improvements in teaching and learning.

The TAG's overall assessment of the feasibility study

The 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 seventeen 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 ICs, 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, seventeen 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 and as reflected in earlier sections of this chapter, 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 U.S., 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.

REFERENCES

- AHELO Consortium (2012b), *AHELO Technical Standards*
[http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/imhe/ahelo/gne\(2012\)16&doclanguage=en](http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/imhe/ahelo/gne(2012)16&doclanguage=en)
- AHELO Consortium (2011), *Test Administration Manual*
[http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=EDU/IMHE/AHELO/GNE\(2011\)21/ANN5/FINAL&doclanguage=en](http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=EDU/IMHE/AHELO/GNE(2011)21/ANN5/FINAL&doclanguage=en)
- AHELO Consortium (2010), *AHELO Assessment Design*
[http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/imhe/ahelo/gne\(2010\)17&doclanguage=en](http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/imhe/ahelo/gne(2010)17&doclanguage=en)
- OECD (2012), *Assessment of Higher Education Learning Outcomes Feasibility Study Report - Volume 1 - Design and Implementation*, Paris
<http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume1.pdf>

NOTES

1. For example, the PISA Technical Advisory Group.
2. Peter Ewell (chair) (USA), Vaneeta D’Andrea (UK), Motohisa Kaneko (Japan), V. Lynn Meek (Australia), Paul Holland (USA), Keith Rust (USA), Frans Van Vught (Netherlands), and Robert Wagenaar (Netherlands).
3. Paul Holland and Keith Rust resigned and were replaced by Stuart Elliott (USA) and Chris Skinner (UK).
4. One or two additional meetings will be scheduled for the purpose of providing a definitive recommendation on the feasibility of AHELO.
5. Note that all these recommendations must be viewed in the context of what was known to the TAG at the time they were made. Some recommendations might have been framed differently had members of the TAG known how certain issues were going to eventually turn out.

ANNEX C – ADDITIONAL TABLES AND FIGURES

Table C1 - Generic skills assessment multiple-choice items item-level non-response by item rotation (n=10657)

Rotation	Skipped		Unreached		Total	
	n	%	n	%	n	%
1	0.7	2.8	2.5	10.0	3.2	12.8
2	0.6	2.4	2.5	10.0	3.1	12.4
3	0.7	2.8	2.7	10.8	3.4	13.6
4	0.7	2.8	2.5	10.0	3.1	12.4

Table C2 - Economics assessment item-level non-response by item group (n=6242)

Task/Module	Skipped		Unreached		Total	
	n	%	n	%	n	%
MCQ1	0.9	7.5	1.7	14.2	2.6	21.7
MCQ2	0.7	5.8	1.9	15.8	2.6	21.7
MCQ3	0.8	6.7	1.8	15.0	2.6	21.7
MCQ4	0.8	6.7	1.8	15.0	2.6	21.7
CRT1	1.2	17.1	0.0	0.0	1.2	17.1
CRT2	0.4	6.7	0.0	0.0	0.4	6.7

Table C3 - Engineering assessment item-level non-response by item group (n=6078)

Task/Module	Skipped		Unreached		Total	
	n	%	n	%	n	%
MCQ1	0.2	4.0	0.5	10.0	0.6	12.0
MCQ2	0.2	4.0	0.5	10.0	0.7	14.0
MCQ3	0.2	4.0	0.5	10.0	0.7	14.0
MCQ4	0.2	4.0	0.4	8.0	0.6	12.0
MCQ5	0.2	4.0	0.4	8.0	0.6	12.0
MCQ6	0.2	4.0	0.6	12.0	0.8	16.0
CRT1	0.1	1.4	0.0	0.0	0.2	2.9
CRT2	0.2	2.5	0.0	0.0	0.2	2.5
CRT3	0.1	1.7	0.0	0.0	0.1	1.7

Table C4 - Generic skills cognitive labs follow-up questions

1.	Did the instructions provide adequate information for you to understand what is expected of you in performing the task? If not, please explain.
2.	Did the question make sense to you? If not, please explain.
3.	How did you decide which items and information to use in answering the question?
4.	What was your strategy for working through the task?
5.	Did you find the performance task engaging? Please explain.

Table C5 - Student feedback collected during focus groups (economics and engineering strands)

	Economics (n=406) Agree or strongly agree (%)		Engineering (n=308) Agree or strongly agree (%)	
	Task 1	Task 2	CRTs	MCQs
There was good linkage between the questions in each task	47.8	48.3	55.2	35.8
The task challenged me to think	78.1	82.2	74.2	62.1
The materials stimulated my interest in the task	46.4	60.1	54.5	46.6
The task was relevant to the content being assessed	49.0	58.0	59.9	61.5
The task made me apply knowledge and skill in real-world ways	39.4	41.2	74.9	40.3
The task covered topics relevant to my program	51.3	61.4	65.2	60.1
The task tested an appropriate range of knowledge and skills	52.1	59.8	53.1	65.6
The task was relevant to my program of study	43.9	54.7	63.6	65.2
The task was relevant to future professional practice	26.0	28.6	54.9	40.0
The task required me to apply capability gained in my program	62.1	70.7	68.8	66.7
The test materials were easy to understand	41.5	45.0	52.2	65.7
The time available was sufficient for me to complete this task	36.7	40.2	28.3	35.8

Table C6 - Overall instrument reliability estimates, by strand by countries

Strands	Countries	Plausible values	Final plausible values (with conditioning)
Generic skills (n=10657)	Country 1	0.66	
	Country 2	0.62	
	Country 3	0.65	
	Country 4	0.69	
	Country 5	0.76	
	Country 6	0.83	
	Country 7	0.83	
	Country 8	0.72	
	Country 9	0.72	
	Total	0.82	0.83
Economics (n=6242) ¹	Country 1	0.82	
	Country 2	0.62	
	Country 3	0.58	
	Country 4	0.73	
	Country 5	0.56	
	Country 6	0.74	
	Total	0.80	0.84
Engineering (n=6078)	Country 1	0.58	
	Country 2	0.65	
	Country 3	0.66	
	Country 4	0.45	
	Country 5	0.62	
	Country 6	0.48	
	Country 7	0.55	
	Country 8	0.49	
	Country 9	0.61	
	Total	0.65	0.75

Table C7 - Generic skills inter-scorer reliability statistics (n=10657)

CRT	Criteria	\bar{X}_d	$s(\bar{X})$	% _A	ρ	κ
1	ARE	0.53	0.60	53.0	0.83	0.35
	PS	0.56	0.60	50.4	0.84	0.34
	WE	0.53	0.61	53.7	0.84	0.37
2	ARE	0.54	0.59	51.7	0.85	0.35
	PS	0.59	0.62	48.8	0.83	0.28
	WE	0.55	0.61	52.2	0.85	0.35

Table C8 - Economics scoring inter-scorer reliability statistics (n=8325)

CRT	Item	\bar{X}_d	$s(\bar{X})$	% _A	ρ	κ
1	1	0.10	0.27	90.74	0.94	0.80
	2	0.16	0.34	84.98	0.87	0.73
	3	0.04	0.16	95.66	0.93	0.88
	4	0.06	0.14	93.50	0.75	0.71
	5	0.08	0.26	92.36	0.87	0.78
	6	0.12	0.29	88.28	0.64	0.53
	7	0.09	0.22	90.64	0.78	0.69
2	1	0.23	0.39	76.68	0.57	0.46
	2	0.27	0.44	75.30	0.72	0.53
	3	0.21	0.38	81.00	0.51	0.42
	4	0.22	0.40	79.68	0.81	0.59
	5	0.21	0.39	80.46	0.72	0.56
	6	0.35	0.49	70.22	0.69	0.51

Table C9 - Engineering inter-scorer reliability statistics (n=8084)

CRT	Item	\bar{X}_d	$s(\bar{X})$	% _A	ρ	κ
1	1	0.24	0.43	75.57	0.56	0.38
	2	0.23	0.39	76.80	0.56	0.45
	4	0.22	0.38	79.36	0.64	0.48
	5	0.24	0.39	76.91	0.60	0.45
	6	0.26	0.36	75.53	0.43	0.37
	7	0.21	0.39	79.80	0.64	0.47
2	2	0.06	0.20	94.26	0.82	0.76
	3	0.21	0.35	81.42	0.69	0.57
	4	0.42	0.56	66.60	0.65	0.44
	6	0.19	0.37	80.57	0.59	0.45
	7	0.15	0.31	85.20	0.63	0.49
	8	0.53	0.65	59.88	0.79	0.43
3	1	0.03	0.12	97.29	0.92	0.87
	2	0.19	0.37	81.07	0.69	0.57
	3	0.09	0.24	91.13	0.88	0.79
	4	0.10	0.25	90.11	0.85	0.77
	5	0.16	0.32	83.63	0.74	0.62
	6	0.54	0.72	61.76	0.67	0.37

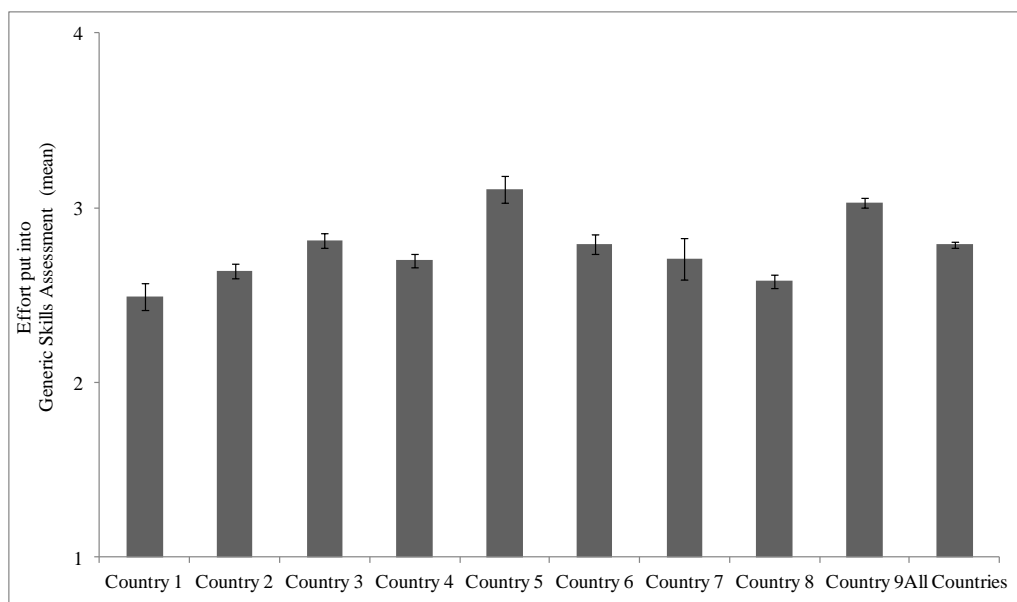
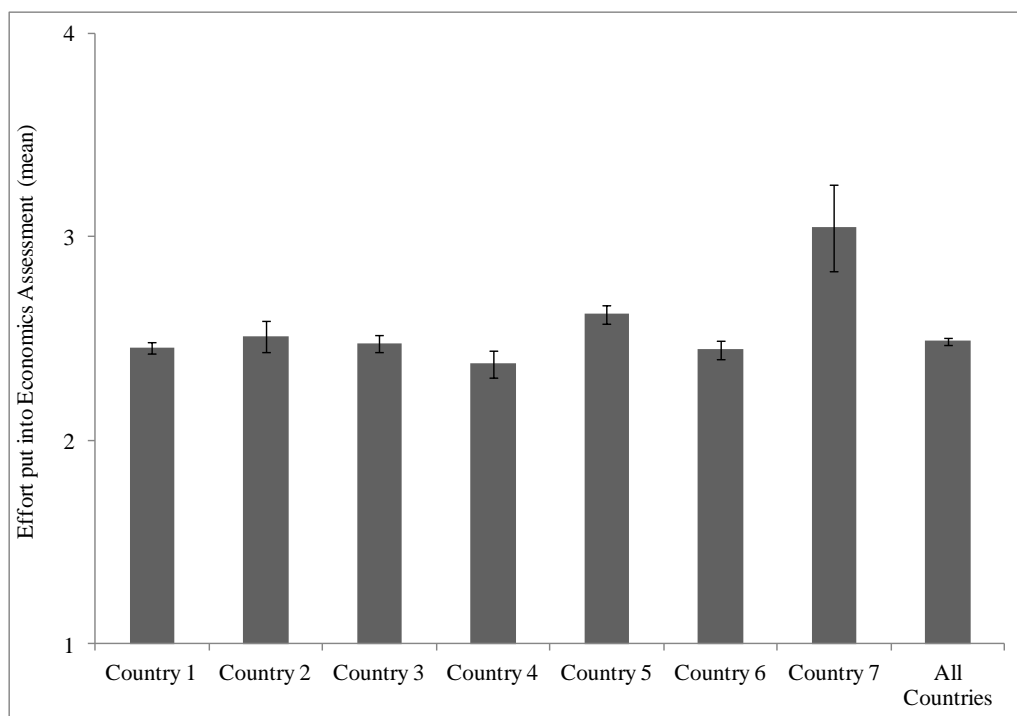
Figure C1 - Self-reported effort put into the generic skills assessment, by country (n=10657)**Figure C2 - Self-reported effort put into the economics assessment, by country (n=6242)**

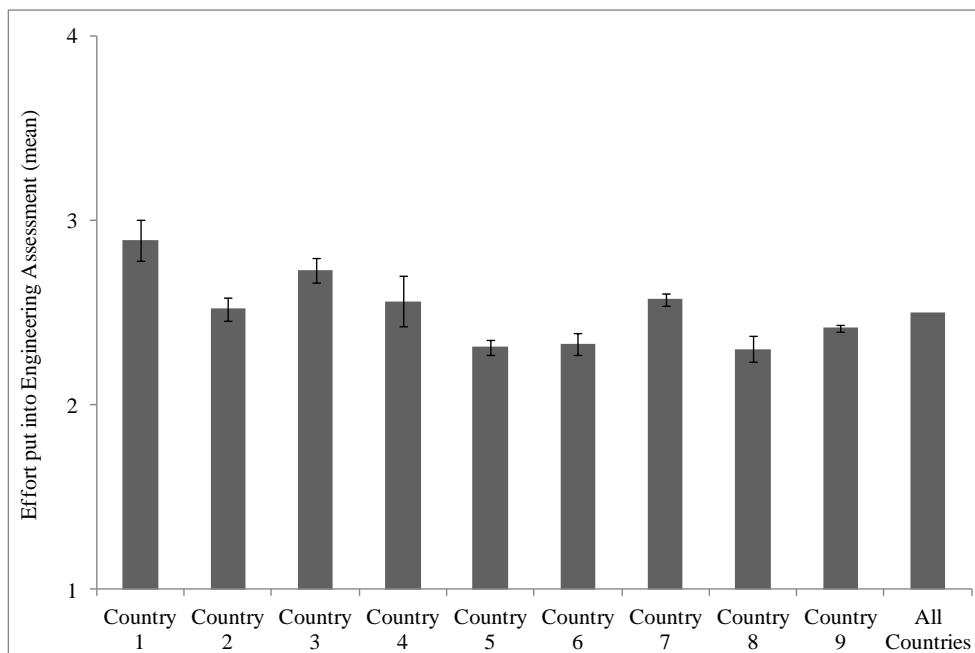
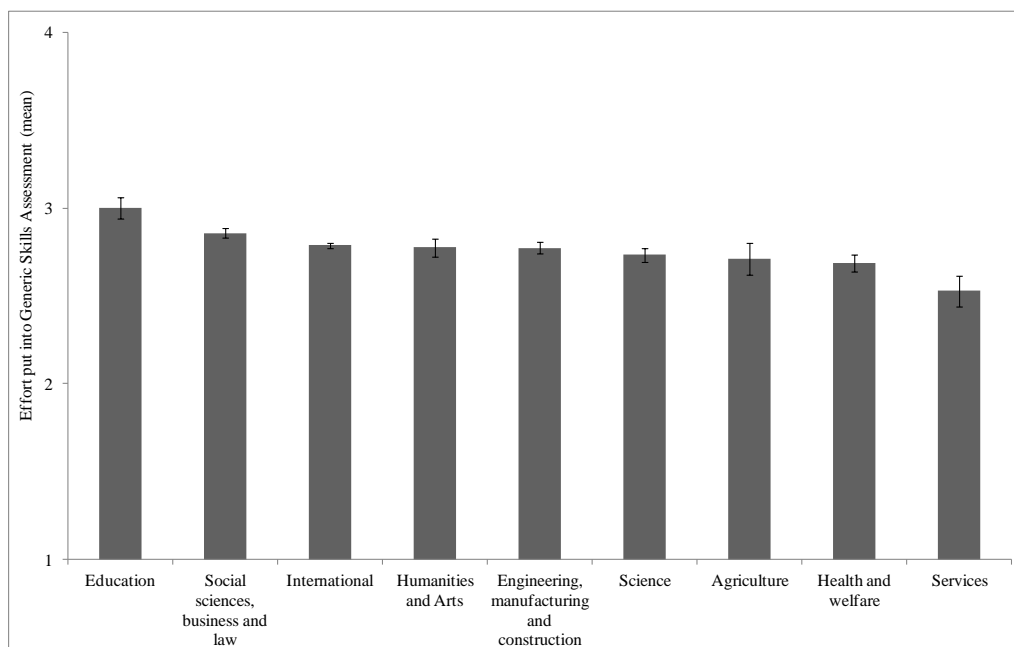
Figure C3 - Self-reported effort put into the engineering assessment, by country (n=6078)**Figure C4 - Self-reported effort put into the generic skills assessment, by field of education (n=10657)**

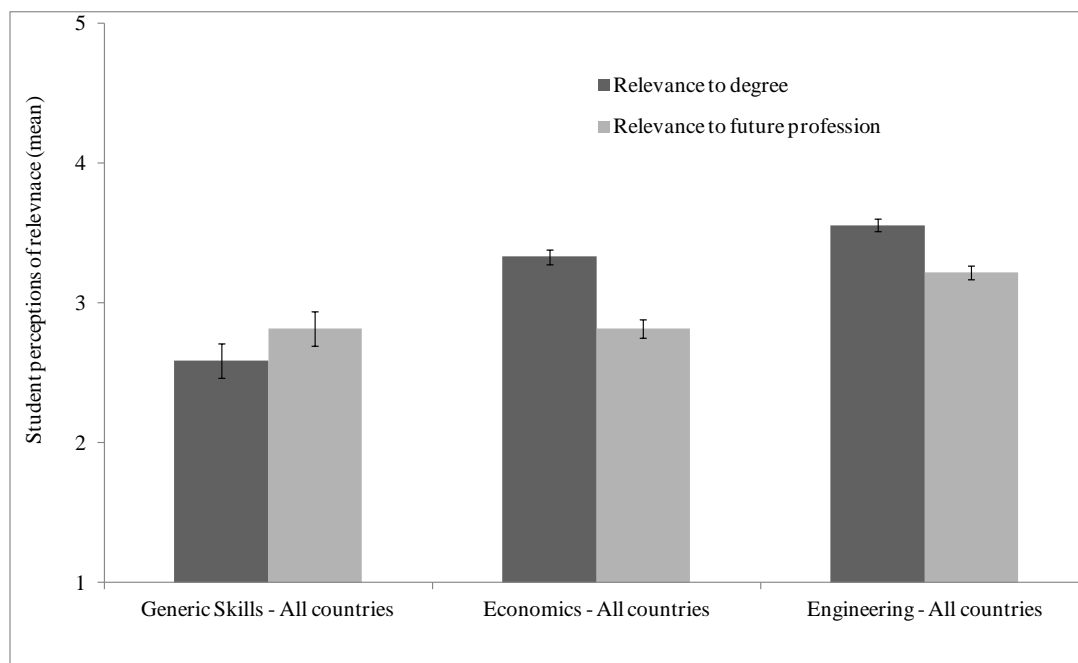
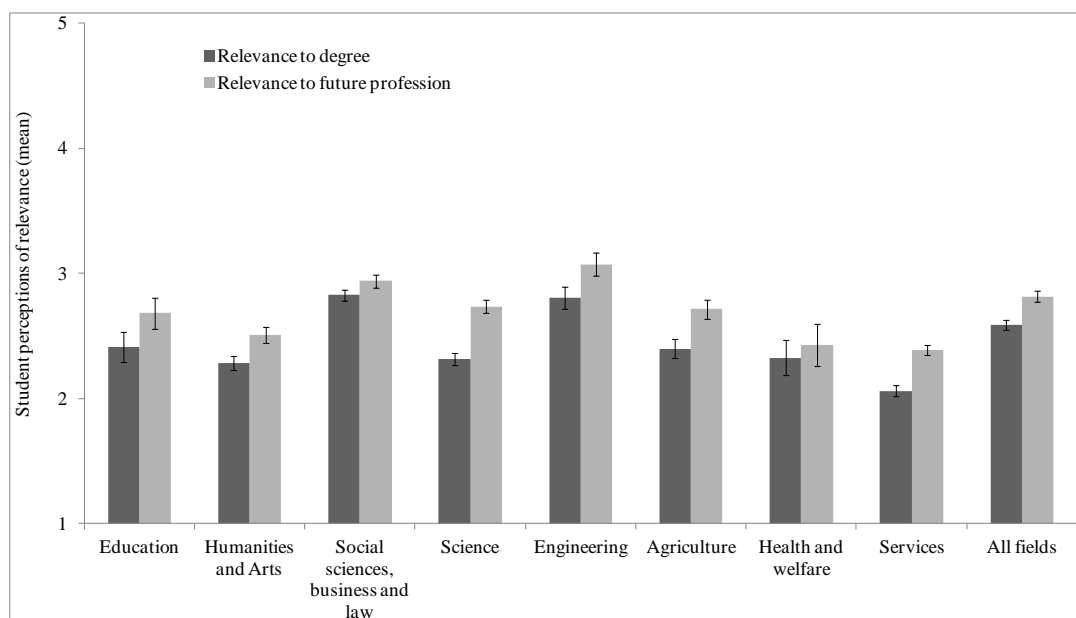
Figure C5 - Student perceptions of relevance of the assessment instrument, by strand**Figure C6 - Student perceptions of relevance of the generic skills assessment, by field of education (n=10657)**

Figure C7 - Generic skills score and self-reported academic performance, by country (n=10657)

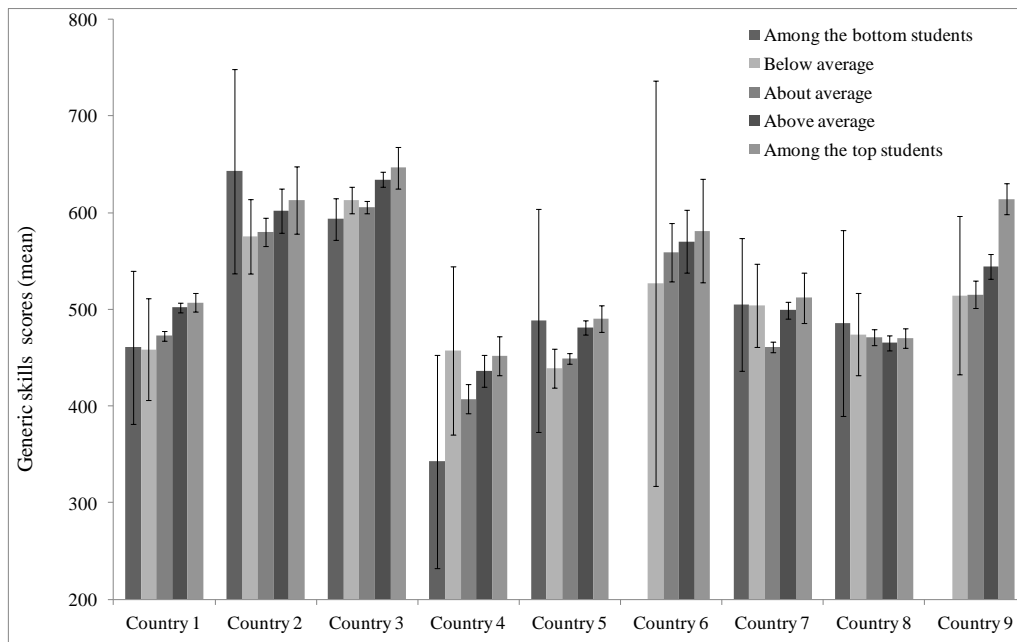


Figure C8 - Economics score and self-reported academic performance, by country (n=6242)

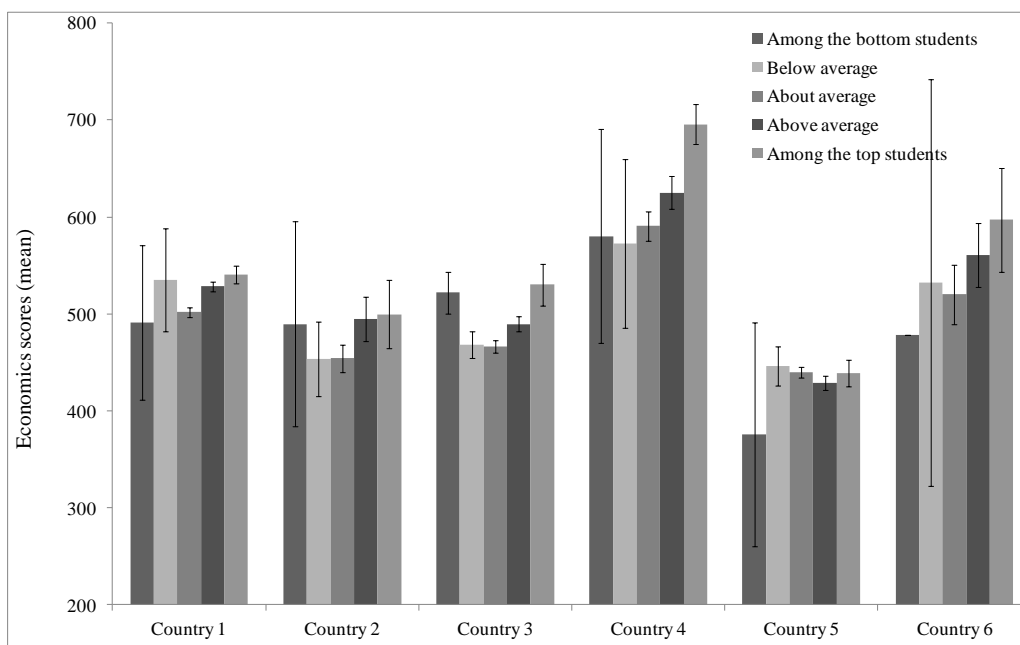


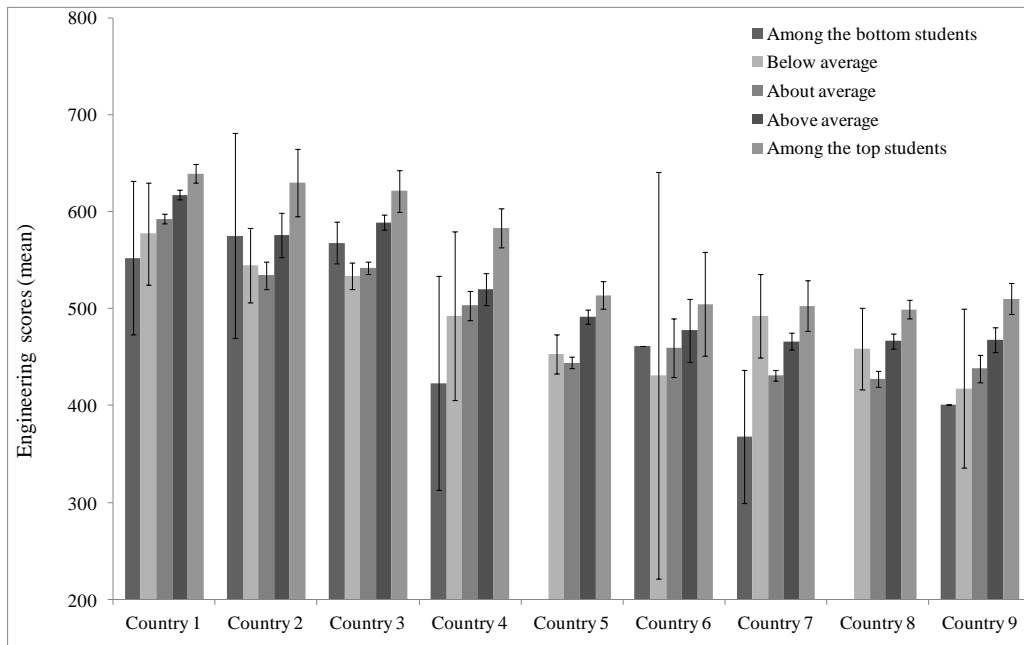
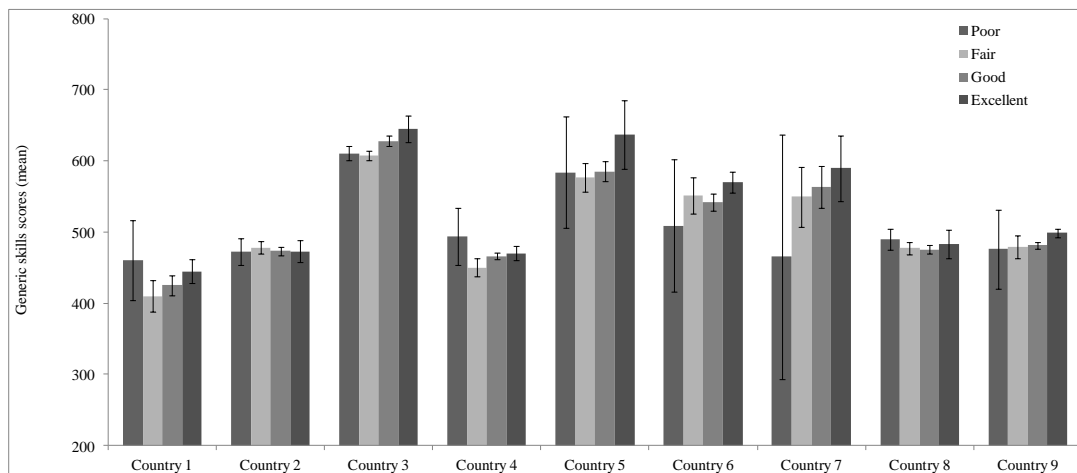
Figure C9 - Engineering Score and self-reported academic performance, by country (n=6078)**Figure C10 - Generic skills scores and overall education satisfaction, by country (n=10657)**

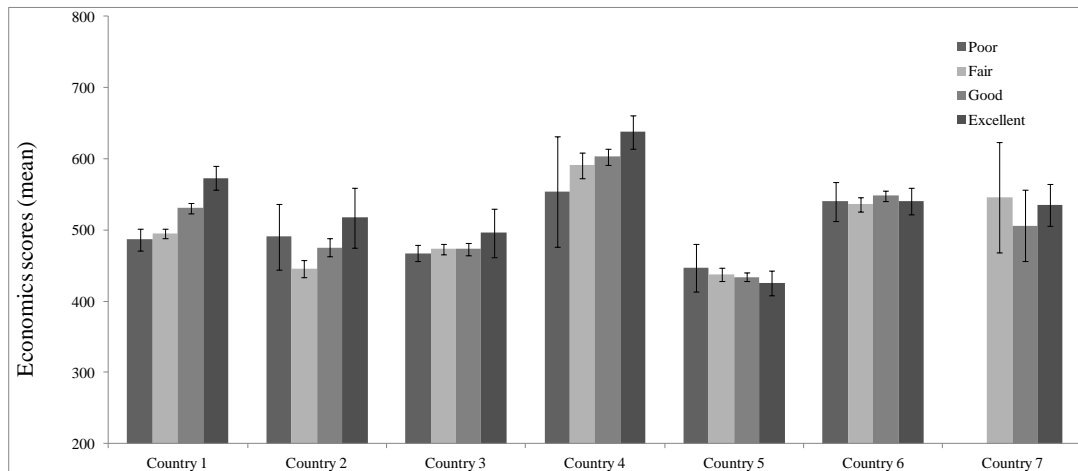
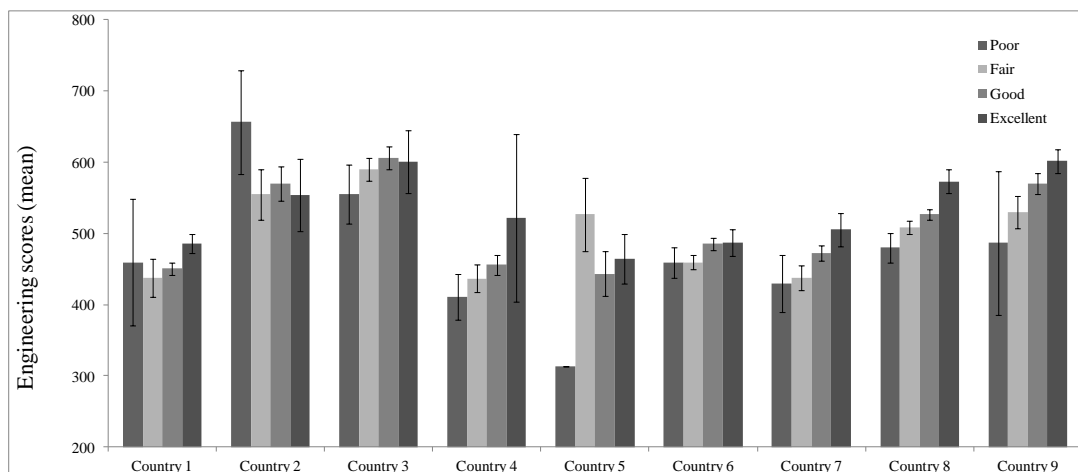
Figure C11 - Economics score and overall education satisfaction, by country (n=6242)**Figure C12 - Engineering score and overall education satisfaction, by country (n=6078)**

Figure C13 - Generic skills assessment variable map (n=10657)

Score	Students	Items
		CR2_PS.5
		CR2_ARE.5
		CR1_PS.5
		CR1_ARE.5 CR1_WE.5 CR2_WE.5
841		MC16
		MC11
		X
		X CR2_ARE1.4 CR2_WE.4
		X MC44 CR1_WE.4 CR2_PS.4
		XX MC10 CR1_ARE.4 CR1_PS.4
		XX MC25
		XX MC27
		XX MC21 MC28
701		XXX MC41 MC49
		XXXX
		XXXX MC20 MC24
		XXXX MC6 MC40
		XXXX MC8 MC35 MC45 MC48 MC53 CR2_ARE.3 CR2_WE.3
		XXXX MC7 MC31 MC43 CR2_PS.3
		XXXXXXXX MC54 CR1_WE.3
		XXXXXX MC18 MC34 MC37 CR1_ARE.3
		XXXXXXXX MC23 MC52 CR1_PS.3
		XXXXXXXX MC4 MC9 MC13 MC32
		XXXXXXXX MC14 MC55
		XXXXXXXX MC5 MC22 MC26
580		XXXXXXXX
		XXXXXXXX MC15 MC39
		XXXXXXXX MC42
		XXXXXXXX MC3 MC30 MC47 CR2_WE.2
		XXXXXXXX MC38
		XXXXXXXX CR2_ARE.2 CR2_PS.2
		XXXXXXXX MC29 MC51 CR1_WE.2
		XXXXXXXX MC46
		XXXXXXXX
		XXXXXXXX MC19
		XXXXXX MC33 CR1_ARE.2 CR1_PS.2
		XXXXXX MC50
449		XXXX MC1
		XXXX
		XXX MC2
		XXX
		XX
		XX CR2_WE.1
		X
		X
		X CR2_PS.1
		X CR2_ARE.1
		X CR1_WE.1
318		
		CR1_PS.1
		CR1_ARE.1

Figure C14 - Economics assessment variable map (n=6242)

Score	Students	Item
1085		CR1G
		CR1D
		CR1F
937	X	CR2C
		CR1C
		MC5 CR2E
		CR1E
		MC18
789	X	CR2D
	X	MC22
	X	MC6 MC28 CR1B
	X	MC9 MC15 CR1A
	X	MC3 MC4
	XX	MC23 MC29 MC31 CR2G
	XX	MC13 MC25 MC38 MC47 CR2A CR2B
	XXX	MC26 MC36
	XXX	MC19 MC27 MC33 MC41
	XXXX	MC32 MC34 MC37 MC40
	XXXXX	
	XXXXXX	MC17 MC21 MC46
	XXXXXXX	MC7 MC11 MC35
641	XXXXXXXX	MC45 MC48
	XXXXXXXX	MC2 MC10 MC20 MC39
	XXXXXXXX	
	XXXXXXXX	MC12
	XXXXXXXXXX	MC42
	XXXXXXXXXX	MC24
	XXXXXXXXXX	MC8 MC43
	XXXXXXXXXX	MC30
	XXXXXXXX	MC1
	XXXXXX	
493	XXXXX	MC44
	XXXX	
	XXX	
	XX	MC14
	XX	
345	X	
	X	

Figure C15 - Engineering assessment variable map (n=6078)

Score	Students	Items
1014		
		CR35
		CR32
		CR31
		CR12
825		MC30
		MC5 MC19 MC26 CR16 CR22 CR23
		MC15 CR15
	X	CR14 CR27
	X	MC17 MC24 CR11
	X	MC12 MC21
	X	MC11
	XX	
636	XXX	MC25 MC27 CR26
	XXX	MC14 CR28
	XXXX	MC2 CR24
	XXXXXX	
	XXXXXXXX	MC23 CR17
	XXXXXXXX	MC4 MC18 CR21 CR36
	XXXXXXXX	
	XXXXXXXX	CR25
	XXXXXXXX	MC6 MC20 MC28
	XXXXXXXX	MC22
448	XXXXXXXX	MC10
	XXXXXXXX	MC1
	XXXXXXXX	MC3 MC8 MC9
	XXXXXXXX	MC13 CR13
	XXXXXXXX	
	XXXXXXXX	MC16
	XXXXXX	MC7
	XXXXXX	MC29
	XXXX	CR33
	XXX	
259	XXX	CR34
	XX	
	X	
	X	
	X	

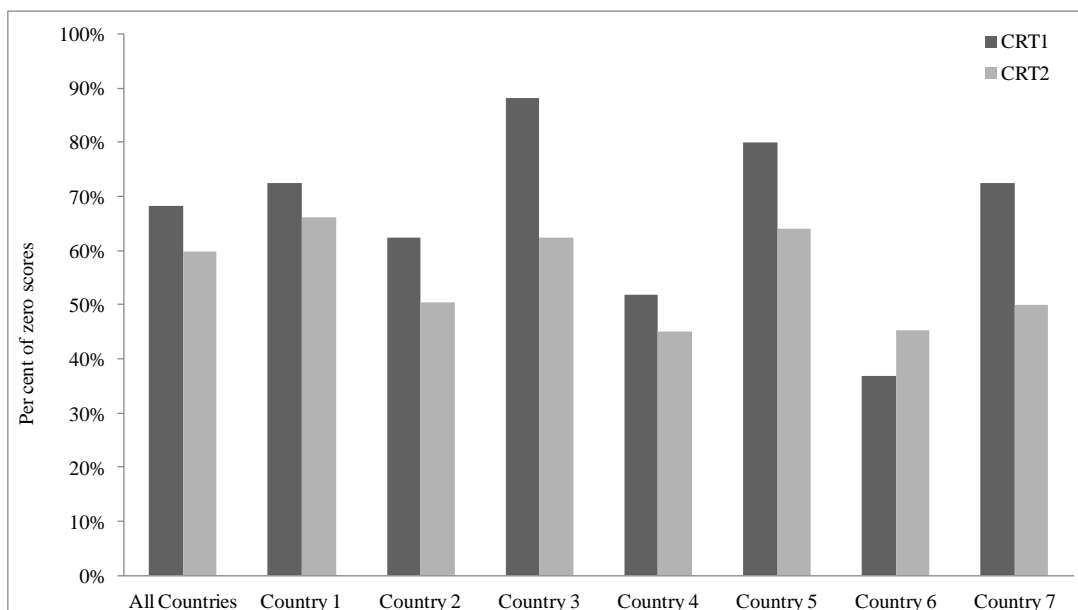
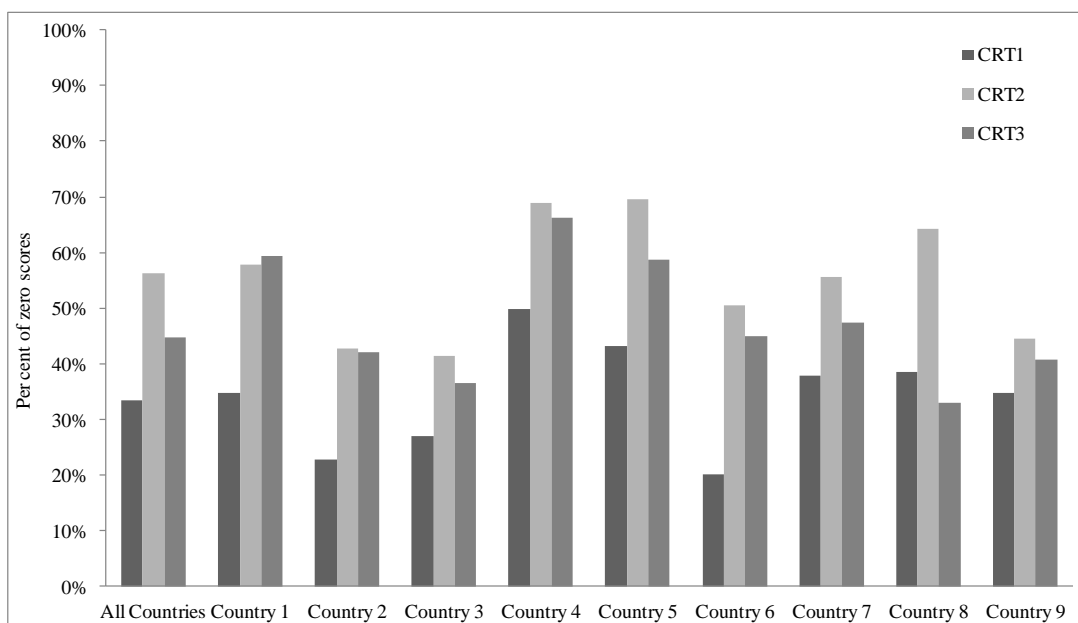
Figure C16 - Economics assessment zero scores, by CRT and country² (n=6242)**Figure C17 - Engineering assessment zero scores, by CRT and country (n=6078)³**

Figure C18 - Generic skills score variance explained by effort, by country and task type
(n=10657)⁴

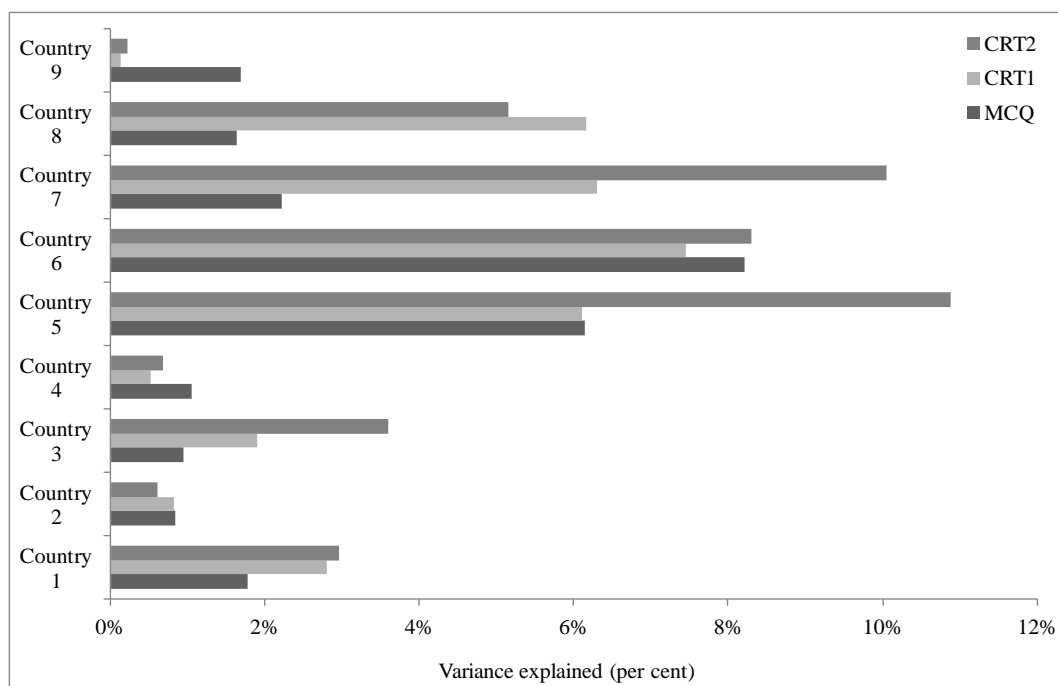


Figure C19 - Economics score variance explained by effort, by country and task type (n=6242)

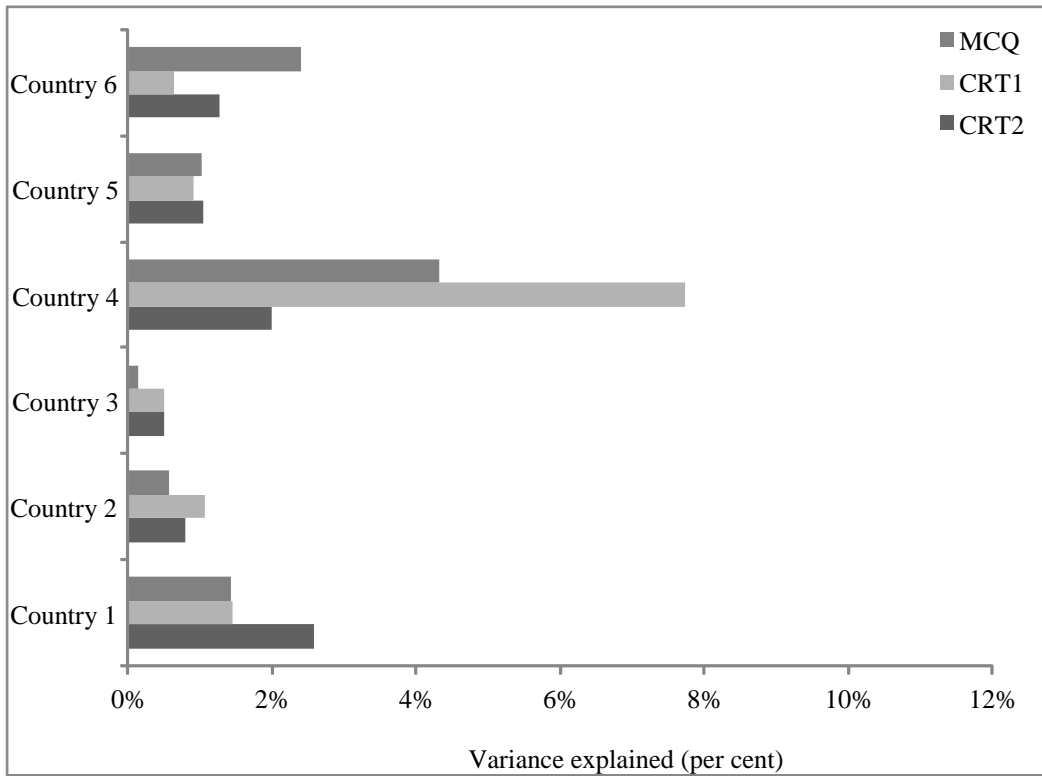
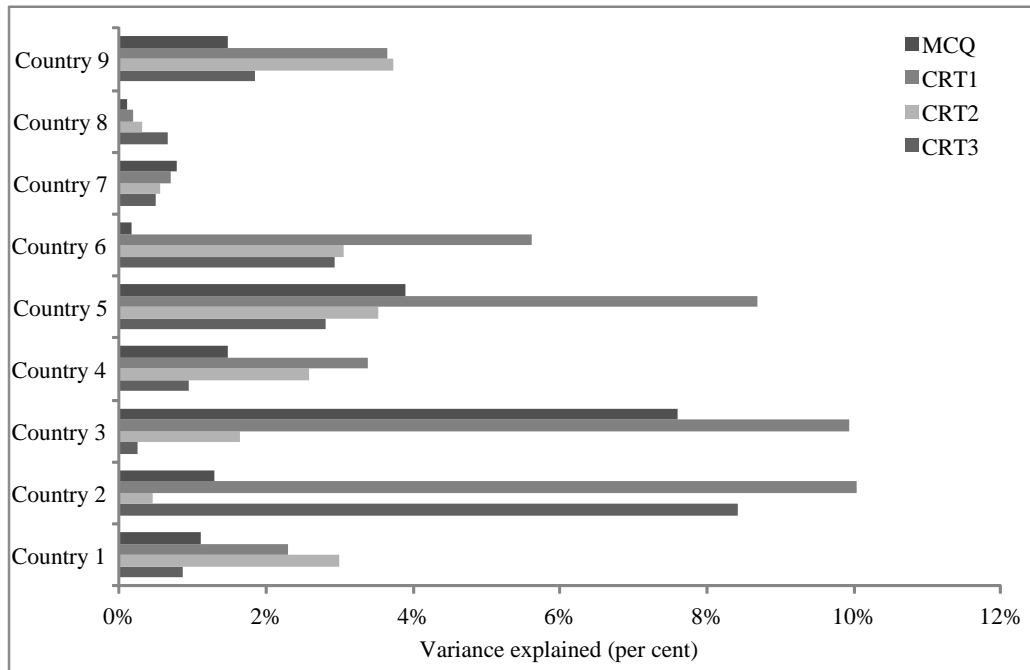


Figure C20 - Engineering score variance explained by effort, by country and task type
(n=6078)⁶



NOTES

- 1 Reliability estimates were not calculated for one country due to its small sample size.
- 2 Confidence interval information not available.
- 3 Confidence interval information not available.
- 4 Confidence interval information not available.
- 5 Confidence interval information not available.
- 6 Confidence interval information not available.

ANNEX D – TAG TERMS OF REFERENCES

[Extract from Document EDU/IMHE/AHELO/GNE(2010)19]

1. The TAG is a consultative group that provides guidance of a technical, scholarly or practical nature.
2. The TAG is managed by the Contractor for Module E.
3. The Contractor for Module E is responsible for suggesting membership. The overriding principle guiding the selection of members for the TAG is relevant expertise. TAG members do not represent specific stakeholder groups or provide policy advice.
4. As part of the contractual agreement between the Module E Contractor and OECD, the Contractor is requested to establish the Technical Advisory Group (TAG) comprised of experts and individuals who have a leading operational role in the AHELO feasibility study.
5. The TAG will be led by a Chair.
6. TAG composition will be approved by the OECD Secretariat in consultation with the AHELO Group of National Experts (GNE). The Contractor will require selected TAG members to sign a confidentiality agreement.
7. The TAG could be consulted on matters such as instrument development, translation and adaptation procedures, validation activities, scoring and verification procedures, or feasibility evaluations.
8. When appropriate, the AHELO GNE shall also seek the advice of the TAG on these or other matters, either directly or through the OECD Secretariat.
9. The TAG will review and provide feedback on documents when requested and its members may participate in meetings organised as part of the AHELO Feasibility Study.
10. The Module E Contractor is also responsible for organising and supporting meetings of the TAG and will be responsible for managing the logistics and bearing the costs of such meetings, including provision of meeting facilities, travel, and the compensation of members of the TAG for face-to-face meetings.
11. The TAG will have at least one face-to-face meeting between July 2010 and June 2011. Further meetings will be conducted via email and by teleconference.
12. TAG communication and meetings will be conducted in English.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD member countries are: Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The European Union takes part in the work of the OECD.

OECD Publishing disseminates widely the results of the Organisation's statistics gathering and research on economic, social and environmental issues, as well as the conventions, guidelines and standards agreed by its members.



www.oecd.org/edu/ahelo

Over the past 5 years, the OECD has carried out a feasibility study to see whether it is practically and scientifically feasible to assess what students in higher education know and can do upon graduation across diverse countries, languages, cultures and institution types. This has involved 249 HEIs across 17 countries and regions joining forces to survey some 4 900 faculties and test some 23 000 students.

This second volume of the feasibility study report presents the data analysis and national experiences.

It follows a first volume on design and implementation which was published in December 2012.

A third volume will be published in April 2013 presenting further insights and the summary of discussions from the AHELO Feasibility Study Conference (taking place in March 2013).

Contents

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

Chapter 8 – National experiences

Chapter 9 – Role of the Technical Advisory Group

More information on www.oecd.org/edu/ahelo

Contact us: ahelo@oecd.org