

ASSESSMENT OF HIGHER EDUCATION  
LEARNING OUTCOMES

**AHELO**

FEASIBILITY STUDY REPORT

**VOLUME 3**

FURTHER INSIGHTS



# **Assessment of Higher Education Learning Outcomes**

## **Feasibility Study Report**

### **Volume 3 – Further Insights**



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](mailto: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](mailto:info@copyright.com) or the Centre français d'exploitation du droit de copie (CFC) at [contact@cfcopies.com](mailto:contact@cfcopies.com).

---

### ACKNOWLEDGEMENTS AND CREDITS

Chapter 10 was written by Henry Braun, the chair of the Value-Added Measurement expert group, in collaboration with the other members of the group (Daniel McCaffrey, Jeffrey Steedle, Julián Mariño, Ou Lydia Liu, Peter Ewell, Richard Arum, Timothy Rodgers, Torbjørn Hægeland, and Diane Lalancette).

Chapter 11 and Annex F were prepared by Cécile Bily, with input from Alenoush Saroyan. Special thanks are also due to Cécile for editing and preparing this report for publication.

We would also like to thank the speakers, panelists and moderators of the AHELO feasibility study conference: Jamie Merisotis, Andreas Schleicher, Peter Ewell, Peter Coaldrake, Jan Levy, Satoko Fukahori, Saana Radi, Fiorella Kostoris, Steve Egan, Harvey Weingarten, Marita Aho, Michael Hoffmann, Nevena Vuksanović, David Robinson, Eva Egron-Polak, Roman Nedela, Kukio Kishimoto, Alfredo Dajer Abimerhi, Deborah Newman, Marta Encinas-Martin, Joanne Caddy, Rebecca Hughes and Eduardo Cascallar.

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, Deborah Roseveare, Richard Yelland, Karine Tremblay, Fabrice Hénard, Valérie Lafon, Anna Glass, Leslie Diamond 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
CHAPTER 10 VALUE-ADDED MEASUREMENT: REPORT FROM THE EXPERT GROUP MEETING9	
Introduction.....	10
Value-Added Analysis and Value-Added Models.....	12
Different approaches to Value-Added Analysis.....	14
Value-Added Analysis in Norwegian in K-12 education.....	14
Value-Added Analysis in Colombian Higher Education Institutions .....	15
Value-Added Analysis in U.S. Higher Education Institutions .....	16
Critical Issues for Value-Added Analyses.....	17
Designs for Value-Added Analysis .....	19
The Longitudinal Design.....	19
The Double Cross-Sectional Design .....	20
The Hybrid Design.....	21
AHELO vs. PIAAC .....	21
Further Discussion Points and Recommendations .....	22
Summary.....	24
Meeting Participants .....	25
Meeting Agenda .....	26
CHAPTER 11 - CONFERENCE REPORTING .....	31
Measuring learning outcomes: what for and for whom?.....	32
The feasibility study: a starting point .....	35
What purposes are attributed to AHELO?.....	35
An evaluation within the wider environment of higher education .....	36
What role should stakeholders have? .....	37
Is there an added value for an international assessment?.....	38
Learning from each other: the experience of the feasibility study participants .....	38
What's in it for us? How to motivate participants?.....	39
What do you assess? .....	41
Types of learning outcomes.....	41
The contextual dimension .....	42
Measuring value added.....	43
How do you measure learning outcomes? .....	43
Types of measures .....	43
CRTs or MCQs.....	44
Timelines, test administration and delivery .....	45

---

Was AHELO worth the cost and should it continue? .....	45
ANNEX E: CONFERENCE AGENDA .....	47
ANNEX F: WORKSHOP EXERCISES.....	51
Workshop 1 - How can measures of learning outcomes provide a valid and valuable response to today's higher education challenges? .....	51
Exercise 1 - Identify Challenges.....	51
Exercise 2 - What are most important learning outcomes in higher education? .....	54
Exercise 3 - The different measures of learning outcomes .....	57
Workshop 2: What are the key challenges in developing an international measure of learning outcomes? .....	65
Exercise 4 – Generic Skills, Discipline Specific Skills or a blended approach .....	65
Exercise 5 – CRTs vs. MCQs.....	76
Exercise 6 – Enhancing student response rates.....	84
Workshop 3: How can we combine an assessment of learning outcomes that is useful to institutions with wider policy goals? .....	91
Exercise 7 - Types of data and Uses for the data .....	91
Exercise 8 .....	99



## 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 was 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 demonstrated what is feasible and what could be feasible, what worked well and what did not, as well as provided lessons and stimulated reflection on how learning outcomes might be most effectively measured in the future.

The outcomes of the feasibility study are 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;
- a **second volume** on data analysis and national experiences which was published in March 2013;
- the feasibility study **Conference** which took place in Paris on 11-12 March 2013; and
- this **third** and final volume on further insights including the report from the Expert Group on the value-added measurement and the Conference proceedings.



## READERS' GUIDE

The chapter numbering follows from the first two volumes of the reports. Therefore this third volume starts with Chapter 10 (Chapters 1 to 6 having been published in the first volume and Chapters 7 to 9 in the second volume).

**Chapter 10** presents the report from the Expert Group on Value-Added Measurement.

**Chapter 11** synthesises the discussions which took place at the Conference.

### **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.

**CHAPTER 10****VALUE-ADDED MEASUREMENT:  
REPORT FROM THE EXPERT GROUP MEETING**

Henry Braun

This report draws on the presentations and discussions that took place during the Expert Group meeting (31 January – 1 February 2013) in Washington, DC. The particular purpose of the meeting was to examine the benefits and challenges of incorporating a Value-Added Analysis into an eventual AHELO main study.

## Introduction

Over the last decade and longer, there has been an increasing global focus on higher education driven, in part, by its importance in human capital development and the link to economic productivity. With the conjunction of increasing investments in higher education and budgetary constraints, it is not surprising that questions related to cost-effectiveness have arisen. Such questions concern: *i)* how the quality of education provided varies by institution overall and by programme within institution; *ii)* the range of differences in learning outcomes for various sub-populations of students; and *iii)* the costs of providing higher education, as well as the losses due to student attrition. In the United States for example, a commission initiated by then-Secretary of Education Margaret Spellings called for greater transparency and accountability in higher education (U.S. Department of Education, 2006). More recently, the Obama Administration echoed similar themes following the 2013 State of the Union.

These developments in higher education are consistent with a general trend to strengthen the monitoring of public services and to focus more directly on driving improvements and increasing cost-effectiveness (Bird et al., 2005). However, as Bird et al. and others point out, performance monitoring and improvement must be carried out thoughtfully and patiently if they are not to result in unintended negative consequences. This point is also evident in research on organizations in different sectors (e.g., public education, sports, and business) that have registered substantial improvements in performance (Hargreaves and Harris, 2010).

A first step in a programme to answer such questions is to devise appropriate measures of quality, with special attention to the cognitive domain: what students know and can do at the end of their programmes of study, and the extent to which their proficiency improved during the course of their enrolment in higher education. Of course each institution has internal measures of quality, usually determined by individual programmes or departments, and manifested in graduation requirements comprising courses to be taken, examination results and, in some cases, performances, exhibitions and the like.

In view of the trend toward globalization of the world's economies, it is natural for institutions to seek to benchmark their performance against other institutions in their own country, as well as those of other countries. A family of assessments, adapted to different languages and cultures (analogous to the approach used in PISA and PIAAC) is one strategy for developing a common metric for outcomes to support such efforts.

The AHELO feasibility study was a project launched to collect evidence on the feasibility of conducting an international assessment of higher education learning outcomes. It comprised both cognitive assessments and contextual survey instruments. A description of the AHELO feasibility study, lessons learned, and the results of the assessments and contextual surveys are reported in two publications (OECD, 2012a, 2013) and will not be reviewed here. As these two volumes were being compiled, the OECD established an "Expert Group" on "Value-added Measurement" to consider the feasibility and utility of conducting value-added analyses on the data that might be generated by an AHELO main study.

A key motivation for convening the Expert Group was to consider strategies to counter the inevitable use of the results of an AHELO main study to construct crude comparisons of institutions/programmes within, and even across, jurisdictions. The problems inherent in interpreting such decontextualized rankings are well-known (Goldstein and Speiglehalter, 1996) and could, on the one hand, discourage institutional participation and, on the other, generate perverse incentives that would distract both participating institutions and policy makers from constructive improvement strategies. These perverse incentives have been observed in the United States in the K-12 public education sector (Hess and Finn, 2007), as well as in higher education in response to the various international rankings that have achieved some prominence in recent years (Espeland and Sauder 2007; Sauder and Espeland 2009). Thus, participating countries and institutions will want to have indicators that are both more credible and more relevant measures of institutional effectiveness than are raw results.

The specific focus of the Expert Group was a statistical approach called value-added analysis (VAA) or value-added measurement. In general, the intention of VAA is to estimate the contributions of individual institutions to the academic progress their students make during their period of enrolment. This is the approach taken by the Expert Group during its deliberations. However, it was pointed out that it is possible to adopt a longer range view of VAA. For example, one could take as a criterion graduates' cumulative earnings over a fixed period (once they enter the workforce) and attempt to estimate institutions'/ programmes' contributions to the observed differentials in earnings (Rodgers, 2007).

Simplifying somewhat, VAA generates adjusted test results by taking into account both differences in the contexts in which institutions operate and differences in their students' prior academic achievements. These adjusted results, appropriately aggregated, are considered to more closely approximate the relative contributions made by different institutions to their students' learning outcomes. To the extent that this approach is successful, the so-called value-added estimates attached to the institutions/programmes may be regarded as a more useful starting point for conversations about improvement, as well as a more acceptable basis for comparison, at least among institutions within a particular jurisdiction. (Although it may not be the purpose of an AHELO main study, in some countries the VAA results could contribute evidence to a more formal higher education accountability system – particularly if all institutions of a particular type participate in AHELO.)

The point is that a student's end-of-programme proficiency level is the result of her entire history of both school-based and out-of-school experiences. As such, her end-of-programme proficiency level is a very noisy indicator of the effectiveness of the present institution. By adjusting for prior achievement and (perhaps) other contextual factors, a VAA attempts to "level the playing field" across institutions that may be serving very different student populations. Consequently, value-added estimates are likely to be better for purposes of benchmarking institutional performance (distinct from selectivity) than are raw scores. That is, institutions with higher value-added estimates are likely to offer better models than institutions with higher raw scores. Although the two groups may overlap somewhat, they are not likely to be identical.

That said, value-added estimates do not directly address the question of whether institutions' students' test scores meet or exceed fixed thresholds. If the thresholds are linked to interpretable outcomes (i.e. are criterion-referenced), then the distributions of student performance on the assessments, administered at the end of their programmes of study, also provide useful information that can be employed by both institutions and policy makers.

This report draws on the presentations and discussions that took place during the Expert Group meeting (31 January – 1 February 2013) in Washington, DC. The particular purpose of the meeting was to examine the benefits and challenges of incorporating a VAA into an eventual AHELO main study. Meeting participants and their institutional affiliations and the meeting agenda are presented in Annex to this Chapter. Meeting participants also benefitted from a literature review of value-added measurement that had been prepared by HoonHo Kim and Diane Lalancette at the OECD (OECD, 2013b).

Some of the questions raised at the outset of the meeting were:

- i. Is conducting some form of VAA feasible and how might the decision to do so influence the design of an eventual AHELO main study?
- ii. Which VAA designs and what value-added models should be examined in advance of a policy decision to introduce a value-added component in an eventual AHELO main study?
- iii. What are some considerations in the trade-offs among different designs and different models?
- iv. Can value-added comparisons be conducted across national boundaries and, if so, with what caveats?
- v. How can indicators based on VAA inform institutional improvement strategies and assist policy makers both in monitoring institutional quality and in resource allocation? In other words, as one participant put it, the ultimate question is: "What works for whom under what circumstances?"

The aim of this report is to provide a synopsis and synthesis of the discussions at the meeting, augmented by suggestions offered by participants subsequent to the meeting. It is organized as follows: The next section provides some background on VAA. This is followed by an enumeration of some critical issues arising in VAA and by a discussion of some different designs for VAA. The report concludes with a brief comparison of AHELO and PIAAC (because PIAAC also targets an adult population and was computer-administered) and an extended discussion of the various issues addressed in the final session of the meeting.

### **Value-Added Analysis and Value-Added Models**

As noted above, VAA is intended to provide estimates of the specific contributions that individual institutions/programmes make to their students' test results, net of other factors associated with achievement. In the public school sector VAA is usually carried out by building a statistical model (value-added model or VAM) that relates a student's current test performance

(the criterion) to some combination of her prior academic achievement, her demographic characteristics and other contextual variables not under the control of the institution. Roughly speaking, the portion of the variation between institutions/programmes in their students' current test results that is not accounted for by the predictors is then attributed to the institutions/programmes.

VAM is sometimes referred to as “growth modelling”. However, this confuses measuring growth, which is a purely descriptive exercise, with an attempt to make a credible causal inference regarding the impact of an institution/programme on student learning. Strictly speaking, the term VAM should be reserved for situations in which at least one of the predictors in the student model is a prior test score, so that there are at least two test scores for each individual. Ideally, the score should be in the same or similar discipline, although this is not always possible. In general, it is beneficial to employ multiple prior test scores from different subjects because they can account for a greater proportion of the variance in the criterion than can a single score. When prior test scores are not available and other predictors are included in the statistical model, the term “contextualized attainment model” (CAM) is sometimes used. Both VAMs and CAMs were addressed at the meeting. For background and more details on VAM in public school settings, see OECD (2008) and National Research Council (2010).

Shavelson (2009) provides a general treatment of outcomes assessment in higher education, including a discussion of VAA. In the higher education context, placing prior and current test scores on a common scale is often not possible, ruling out in many cases the measurement of growth as it is usually understood—that is, a pre-test followed by a post-test. In fact, a true VAM is rarely implemented in higher education because of the difficulties in conducting the longitudinal study that is required. Consequently, alternative “short-cut” procedures have generally been employed. These issues are discussed in the following sections. To avoid confusion, we will use VAA as a generic term to refer to the full range of procedures that have been used to arrive at an estimate of an institution's (relative) contribution to student learning.

The statistical models used for VAA differ with respect to the form of the regression model, the stochastic assumptions made about the parameters of the model and the type and extent of the predictors included in the model. With VAM, for example, some models only employ measures of prior academic achievement while others employ a broader set of predictors. It is important to keep in mind that the proper interpretation of the output of a VAM depends both on the statistical characteristics of the VAM and the nature of the predictors. By way of illustration, consider a VAM that includes as predictors students' gender and region of origin, along with prior academic achievement. Then a student's contribution to the value-added estimate is (approximately) the difference between her current score and the expected score for students with a similar profile. Critics note that allowing expectations to be determined, in part, by students' demographic characteristics can serve to perpetuate historic inequalities. On the other hand, not making such adjustments, some argue, results in institutions that enrol substantial numbers of students from traditionally low-achieving populations being inappropriately disadvantaged in cross-institutional comparisons.

At first blush, it would seem that once the decision has been made to use predictors other than prior test scores, then more is better. However, the literature on VAM notes that the inclusion of certain predictors in the regression model can actually result in an over-adjustment. For example, family income will be correlated with true institutional effectiveness if students from better-off families tend to enrol in higher quality institutions. Thus, including family income as a predictor in the model will absorb some of the variation in the criterion due to these true differences, resulting in a range of estimates that is smaller than the range of true differences. This diminishment may well be offset by the inclusion of other predictors and the unavoidable errors of estimation. However, the extent and direction of the final bias is very difficult to determine, which suggests caution in the interpretation of results.

### ***Different approaches to Value-Added Analysis***

During the meeting, various approaches to conducting a VAA were examined with a focus on their advantages and disadvantages in higher education and international comparison settings. In addition, there was discussion of design considerations for both the assessment of generic skills and of discipline-specific skills. The discussion was enhanced by the presentation of T. Haegeland on research findings from Norway regarding VAA in the public school sector and J. Marino on the experience in Colombia of evaluating higher education programmes and the preparations for conducting VAA in the near future.

In the United States, rankings of four-year colleges are usually produced by the media and are based on weighted averages of a number of different indicators (e.g. U.S. News and World Report, 2012), rather than on tested outcomes. In the United Kingdom, various sorts of league tables based on test results, as well as a crude form of VAA have been published by various newspapers (T. Rodgers, personal communication).

At this point, there is a fair amount of experience in the United States and in England with the application of VAA to the evaluation of elementary and secondary schools. Some countries, such as the Netherlands, Norway and Poland employ VAA to a limited extent, while other countries are conducting research on VAA. In the U.S. there is still considerable controversy regarding the proper role of both standardized testing, as well as VAA, in K-12 school and teacher accountability (Braun, 2005). Concerns center mainly on:

- i. the extent to which schools serving particular populations of students may be advantaged or disadvantaged by a VAA;
- ii. the dependence of the results on the particular criterion test used; and
- iii. the magnitude of the uncertainty attached to value-added estimates and the year-to-year volatility of those estimates.

### ***Value-Added Analysis in Norwegian in K-12 education***

Statistics Norway has investigated different VAMs with regard to the variance, bias, and stability over time of the estimates they produce (Haegeland, 2011). One question of particular interest was the impact of using different sets of predictors in estimating schools' value-added. The general finding was that having just a few measures of prior academic achievement is

superior to having (only) an extensive set of student-level characteristics comprising both demographics and family circumstances. In fact, with sufficient indicators of prior academic achievement in the model, there is a minimal incremental contribution of student and family characteristics to the proportion of criterion variance accounted for. This is consistent with results obtained in the U.S. (Ballou, Sanders, and Wright, 2004; Goldhaber, Walch and Gabele, 2013) and is a useful finding because the Norwegian system of administrative register data contains an enormous amount of individual level data, much more than could reasonably be obtained through an AHELO survey. Consequently, having good data on prior academic achievement would be all the more critical for AHELO. With respect to how a proposed VAA should influence the design of the cognitive instruments in AHELO, the conclusions from Norway are very much in line with what has been learned in the U.S.; namely, that the assessment should measure valued learning goals over a broad range of performance levels with measurement error kept as uniform as possible. One design implication is that the pre-test measure(s) should be highly correlated with the post-test (criterion) score.

#### ***Value-Added Analysis in Colombian Higher Education Institutions***

Colombia has had extensive experience in the evaluation of higher education institutions. Since 2003, all students nearing the end of their tertiary studies at all institutions have been required to sit for assessments that have been centrally prepared and scored. Results are made public at the institutional level. However, for analytic purposes, the most useful level of analysis is at the programme or department level. So, for example, one might compare institutions with respect to their mathematics programmes, their engineering programmes, or their history programmes, and so on. In Colombia, all college students have taken a set of eight curriculum-based exams at the end of secondary education. These can be used as measures of prior academic achievement.

In 2009, Colombia instituted a new evaluation system that tests all graduating seniors on a set of general skills comprising critical reading, quantitative methods, writing (essay), citizenship skills and English. In addition, they must sit for an assessment of “specific competencies” that is appropriate for their programme of study. Research studies have been carried out to explore different approaches to conducting VAA, with the goal of having VAA become part of a formal evaluation by the 2013-14 academic year.

An early study (Saavedra and Saavedra, 2010) used scores on a Spanish translation of the Graduate Skills Assessment Test as the criterion and carried out a VAA on a sample of 2000 seniors drawn from 17 higher education institutions. Subsequently, Domingue (2012) conducted a VAA on a sample of more than 50 000 students using scores on the new quantitative methods test as the criterion with the end of secondary education math score and student characteristics as predictors. The findings were informative with respect to model choice. However, value-added estimates were highly correlated with raw scores, suggesting that the predictor set was not accounting for much criterion variance. However, further analysis and, more importantly, replication studies are needed to provide credible evidence for the design of any future VAA component of AHELO.



***Value-Added Analysis in U.S. Higher Education Institutions***

In the U.S., there has been some empirical research on VAA for higher education (Liu, 2008; Liu, 2011; Steedle, 2012). From a methodological perspective, both authors compare the operating characteristics of different approaches to VAA and conclude that in some circumstances the models generate estimates that can reliably distinguish differences in effectiveness among institution/programmes at the extremes of the distribution of effectiveness. At the same time, they offer cautions with respect to interpretation of the estimates and, especially, their use for purposes of high-stakes institutional accountability. Note that in the U.S. most students enrolled in four-year institutions take at least one of two standardized assessments (SAT or ACT) that can be used for adjustment at the individual or group level.

At the meeting there was some discussion of the way that VAA is now being used in the U.S. as one component of an improvement strategy. Reference was made to the Voluntary System of Accountability, an approach to institutional transparency and inter-institutional comparability to which many four-year public institutions have subscribed. Each institution produces a “college portrait” that contains information on various aspects of the institution (e.g. costs, graduation rates, student learning outcomes, etc.) developed according to agreed-upon methodologies and displayed in a common, accessible format. Many institutions are now in the process of piloting a VAA using students’ test performance or some other evidence of student learning as a criterion. The intention is that, eventually, comparable VAA estimates will become a part of institutions’ college portraits. However, there is evidence that such comparability is compromised if different tests are used as the criterion (Steedle, Kugelmass, & Nemeth, 2010).

There are a number of assessments that are being used in the U.S. to generate college-level criterion scores. Among them are the ACT Collegiate Assessment of Academic Proficiency (CAAP), the ETS Proficiency Profile (EPP) and the Collegiate Learning Assessment (CLA). A modified version of the CLA was used in the AHELO Feasibility Study as part of a measure of general skills. (The measure of general skills also included multiple-choice items from the Australian Council for Educational Research assessing reading, writing, and quantitative literacy.) The CLA comprises a set of performance exercises that are intended to assess higher order skills (Klein et al., 2007).

Many institutions, both public and private, are using the CAAP, the EPP (or an abbreviated version of the EPP), or the CLA as a criterion measure. The manner in which the data are employed to inform institutional improvement varies with the criterion measure and the institution. At the meeting it was noted that members of the Council of Independent Colleges are using both CLA scores and the corresponding VAA to inform institutional improvement efforts (Council of Independent Colleges, 2008). Kalamazoo College was cited as a particularly outstanding example (Sutherland et al., 2008). The University of Texas system has also been using the results of a VAA to drive changes in the undergraduate curriculum (Reyes and Rincon, 2008).

### **Critical Issues for Value-Added Analyses**

A number of issues arise within the context of the AHELO and are addressed in Volume 1 of the Feasibility Study Report. They include:

- i. the choice of the constructs or skills to serve as outcomes,
- ii. the construct validity and comparability of the assessment across multiple institutions, languages and cultures,
- iii. the relevance of the assessment to the range of institutions and programmes participating in AHELO,
- iv. the composition of the contextual surveys, and
- v. the nature and comparability of the student samples sitting for the assessment.

Clearly, these are relevant to the utility of the VAA but will be discussed here only if the VAA raises additional concerns. The Expert Group addressed a number of key technical and practical issues related to the possible introduction of VAA in an AHELO main study. Nine of the issues are presented below.

#### **1. Study design**

There are a number of possible study designs including cross-sectional, longitudinal and hybrid designs. Each has advantages and disadvantages with respect to the type and credibility of the information generated, feasibility, timing and cost. These are discussed in a subsequent section.

#### **2. Timing of the criterion assessment**

An obvious complication is that the nominal expected number of years or semesters to obtaining a first degree can differ both within and across national boundaries. Further, the actual distribution of years/semesters to completion can vary substantially. Thus, it is essential to achieve consensus on how to define comparability in “educational exposure” across institutions. For example, one might include students who are within two semesters of graduation and who have completed at least some minimum number of semesters. Alternatively, for the assessment of general skills, it may be valuable to test students after two, three or four semesters of enrolment, as well as when they are near the end of their programme of study.

#### **3. Choice of prior measures**

VAA requires at least one measure of prior academic achievement. In some countries, one or more end-of-secondary education test results can play this role. In other countries, a special assessment would have to be administered to incoming first-year students. This could be the same general skills assessment that is given later in their college careers or it could be an entirely different assessment—for example, one that is modelled on the assessments administered in PISA or PIAAC. Note that if VAA results are to be comparable across countries then a common measure of prior achievement will be necessary.

#### 4. Contextual data

The information culled from the contextual survey plays an especially important role in VAA. Consequently, the nature and extent of the information sought should be guided in part by empirical findings on what factors are strongly associated with student performance. In addition, school characteristics such as degree of selectivity, graduation rates and the like should be obtained as they can be used to identify classes of comparable institutions, at least within a particular jurisdiction. It was pointed out at the meeting that the contextual variables that would be included in a VAM not only would likely differ from one jurisdiction to another, but also their explanatory power would differ as well – and this would greatly complicate the task of making credible comparisons across jurisdictions.

#### 5. Comparison groups

One of the goals of VAA is to make comparisons among institutions/programmes fairer – and more informative – by adjusting for differences in the populations of enrolled students. That said, other considerations must play a role in delineating a collection of comparable units. Overall comparisons (i.e. at the institutional level) are of limited value because of the variation in the effectiveness of different programmes within an institution. As noted above, in Colombia, institutions were compared at the level of programme (e.g. mathematics, engineering, history, etc.). This seems quite reasonable within a country. Again, cross-country comparisons may be more problematic because programmes of study for a particular discipline, especially in the humanities and social sciences, may vary considerably in content from country to country. A different problem arises in smaller countries with relatively few institutions. In that case, within-country benchmarking may not be very useful and a grouping of countries based on geographical proximity and similarity of culture/language could be a reasonable alternative.

#### 6. Student sample

Here the choice concerns whether to administer the assessment to a census, random or convenience sample. The choice depends on the design (i.e. longitudinal, double cross-sectional, hybrid), the nature of the assessment (i.e. generic or discipline-specific outcomes), institutional characteristics and, of course, costs. When a convenience sample is obtained, for example, by asking students to volunteer to sit for the assessment, questions of bias and generalisability naturally arise. In the case of longitudinal designs, student attrition over time raises some of the same questions. In a modified convenience sample, there is an attempt to select students so that the group sitting for the assessment are roughly representative of the student body, but do not constitute a random sample. To the extent that the attempt is successful, the results may be more credible – though still subject to the caveats above.

#### 7. Test-taking motivation

Differential student motivation in taking low-stakes assessments can substantially influence assessment results (Braun, Kirsch, & Yamamoto, 2011; Liu, Bridgeman, & Adler, 2012). Liu et al. randomly assigned college students to three motivational conditions in taking the EPP multiple-choice test and an essay test. They found that conclusions regarding value-added contributions

were strongly related to both students' levels of motivation and item format (multiple-choice vs. essay). With respect to a possible AHELO generally, and a VAA component specifically, it would be essential to employ strategies to motivate students to exert maximal effort and to attempt to achieve some degree of comparability in the motivational strategies implemented across institutions.

#### 8. Instrument design

The credibility and utility of AHELO, overall and with respect to a VAA component, depends on assessing a sufficiently broad range of valued skills and knowledge. Accordingly, the instrument should comprise challenging probes that require extended responses as well as probes that call for shorter responses. Inclusion of forced choice questions is also possible. In any event, the tension between full domain coverage (construct representation) and the need to limit testing burden and costs must be resolved by compromise. One strategy to maximise domain coverage is to use matrix sampling, so that one individual is only exposed to a fraction of the item pool. The disadvantage is that this strategy complicates longitudinal analyses needed for VAM by sacrificing simple comparability.

#### 9. Models

As noted earlier, there are different families of regression models that can and have been used for the VAM approach to VAA. With regard to VAM, model families differ with respect to a number of characteristics, including numbers of levels, use of fixed or random effects, and range of predictors. With regard to other approaches, families differ with respect to the type of adjustment, the use of single level or multi-level models, and so on. The choice of family should depend on findings in the literature, empirical investigations using AHELO feasibility study results, as well as logistical considerations. Further details are provided in the literature review of value-added measurement (OECD, 2013b) and in Steedle (2012).

### **Designs for Value-Added Analysis**

#### ***The Longitudinal Design***

With this design, students are assessed at two or more points in time. There are several variations of this design depending on the timing and nature of the assessments: in one variant, students are tested at the beginning of their first year and toward the end of their programme of study with the same or a psychometrically parallel assessment. This is appropriate for assessing general skills. An alternative is to use for the initial assessment a common set of assessments taken at the end of secondary education, as is the case in Colombia. Of course, various combinations and extensions are possible, including having students also assessed at intermediate points during their studies. This has the advantage of facilitating a more fine-grained analysis of institution/programme effectiveness or allowing for comparing institutions after a common length of exposure for students. Carrying out a VAA for discipline-specific skills is more problematic since entering students cannot be expected to have any more than rudimentary knowledge/skills for many disciplines. This makes administering parallel assessments unfeasible. Instead, what would be called for is a common, initial assessment of

higher order skills in several domains (at a level appropriate for students at the start of tertiary education), selected in part to be highly predictive of success in discipline-specific studies, with the final assessment being discipline-specific.

Although at first blush the longitudinal design appears to be the ideal approach, practical considerations reveal its weaknesses. Most obviously, it requires waiting three to four years for a cohort to complete its programme of study. Since the lengths of programmes of study also vary by country, co-ordinating an international assessment could be very challenging.

Of equally concern is the problem of tracking students and student attrition. Longitudinal data collection requires the same students be tested twice which requires maintaining records to allow for tracking the student. In addition, some students will leave the institution where they were first tested. In some institutions/programmes attrition rates can be both considerable and informative, so that the students tested at the later time point constitute a highly non-random sample of the starting cohort. In that case, the usual interpretation of the value-added estimates must be modified: They now represent the contributions of the units of analysis to student learning for those students who complete the programme. In biostatistics, this is known as an estimate of the “effect of the treatment on the treated”, which is contrasted with an estimate of “intention to treat” (Lachin, 2000).

As do all designs, the longitudinal design raises issues of sampling. If the focus is on general skills, a stratified random sample of entering students may be sufficient to yield credible results. The stratification criteria and the size of the sample will depend on, among other factors, the institutional structure and the expected attrition rate. If the focus is on discipline-specific outcomes, typically all students near completion of the programme would sit for the assessment. Even so, for some programmes it may be necessary to aggregate data over two or more cohorts to yield estimates with sufficient precision. Note that this assumes that all students will have initial assessment results – which may not be the case if only a sample of first year students have been tested and there are no common end-of-secondary education results available.

### ***The Double Cross-Sectional Design***

With this design, a census or representative samples of both first-year and completing students are tested during the same academic year. Then a statistical adjustment is made to account for differences between the two cohorts in the prior academic measures and a “pseudo value-added estimate” is then computed. The statistical adjustment is based on a regression analysis at the unit (i.e. institution/program) level. Further technical details can be found in OECD (2013b) and Steedle (2011).

The obvious advantage of this design is that it does not involve a multi-year follow-up to obtain the desired estimates. On the other hand, stronger assumptions about the appropriateness of the statistical adjustment are needed to justify the credibility of the resulting estimates. In the U.S., where this approach has been employed, concerns have been raised about this design (Banta and Pike, 2007; Shermis, 2008). These concerns include both the use of college admissions test scores (e.g. SAT scores) as a basis for adjustment and the assumption of a linear relationship between the criterion and the baseline assessment.

### ***The Hybrid Design***

This design has features of both the longitudinal and the double cross-sectional designs. For purposes of discussion, assume that programmes of study are four years in length and that census samples are employed. In one version discussed by the Expert Group, at the start of the academic year, both first year and third year students are assessed. Then, approximately 20 months later, both samples are tested again—the younger cohort near the end of their second year and the older cohort near the end of their fourth year. With the implementation of this design, it is possible to conduct an analysis similar to the double cross-sectional using the first year and fourth year results. Moreover, it is also possible to conduct longitudinal analyses for both cohorts and, by linking the two cohorts, to carry out an approximation to a full, four year longitudinal design.

Of course, there are many variations of this design. One of particular interest would have the younger cohort taking the general, higher-order skills assessment on both occasions, while the older cohort takes the general assessment on the first occasion and both the general assessment and the discipline-specific assessment on the second occasion. If sitting for two assessments is considered too burdensome, then the older cohort could be divided into two stratified random samples, with each (approximate) half-sample taking one of the assessments on the second occasion.

It is important to note that in conjunction with the variation among countries in programme lengths, as well as the seasonal differences in academic calendars between the Northern and Southern hemispheres, the implementation of this design would require AHELO testing to extend over approximately two and a half years. Although this is shorter than the time required for a true longitudinal design, it is still longer than the testing windows of other international assessments and, hence, would necessitate major changes in logistics, technical analyses, reporting schedules, and costs.

### **AHELO vs. PIAAC**

The Expert Group devoted some time to comparing and contrasting AHELO and PIAAC, as they are both computer-delivered and there is an overlap in the target populations. The focus of the latter is both on generic skills and labour market outcomes, with the reporting unit typically being a political jurisdiction (e.g. a country, state or province). Comparisons among units on average skill levels, as well as other outcomes are made overall, by age-cohort, or by some other characteristic. Such comparisons are expected and intended. Utilizing the contextual information to attempt to explain why certain jurisdictions achieve superior results and/or greater progress by age-cohort requires substantial secondary analysis.

By contrast, the primary focus with AHELO is on end-of-programme skill levels. Comparisons will likely be made at the institutional and individual programme (within institution) levels. Accordingly comparisons of the distributions of outcomes, as well as value-added estimates based on those outcomes are of equal interest. Incorporating institutional and or programme characteristics offers a direct approach to generating hypotheses about factors associated with greater institutional effectiveness.

There was some discussion regarding the choice of baseline assessments. In addition to the options already mentioned, the possibility of using modified forms of the PIAAC assessment or the PISA assessments were raised. Although linking AHELO to these other surveys could have some benefits, there were concerns about potential ceiling effects of using these two assessments for the AHELO target population.

### **Further Discussion Points and Recommendations**

At the outset of the Expert Group meeting, it was made clear that no decision has been made regarding an AHELO main study and that, should there be a decision to proceed, many questions – political, financial, logistical and technical – would have to be resolved. The Expert Group certainly recognized the many challenges involved in mounting an AHELO main study and that a VAA component would add to those challenges. Nonetheless, done well, the results of a VAA could make a significant contribution to achieving the goals of AHELO with respect to both institutional improvement and accountability. Consequently, the Expert Group was generally supportive of giving serious consideration to incorporating VAA into an AHELO main study. At the same time, it cautioned that a number of issues would have to be addressed before a final decision could be made. Some of the discussion points in that regard are summarized below. They respond to, and extend beyond, the five questions listed in the introductory section of the report.

#### **1. Feasibility**

The Expert Group agreed that implementing a VAA was feasible in the context of an AHELO main study, provided that those responsible for the management of the study and the participating institutions were prepared to discharge the obligations demanded. The logistics of administration were of particular concern and are discussed below.

#### **2. Assessment design**

For general, higher-order skills, the assessment framework should, to the extent possible, maximise domain coverage especially with respect to valued skills. In particular, the assessment should have a “high ceiling” so that measured gains above expectation are not artificially limited. Similarly, for discipline-specific skills, priority should be given to those disciplines for which there is a generally agreed-to body of knowledge and an achievable consensus regarding mastery. In both cases, utilization of different item types, including variants of both multiple-choice and constructed-response items will be required. Note that the validity/feasibility trade-off is affected by whether some response formats can be scored automatically by expert systems. (This is an issue that likely would be addressed in the planning for an AHELO main study.)

#### **3. Administration**

Different designs call for different assessment schedules that must be co-ordinated with both the academic schedules of the various institutions and the formal programme lengths. In addition, students may be on a “fast-track” or a “slow-track” in relation to typical programme completion. Thus, it will be necessary to develop a set of “business rules” that determine when

in their programme of study students should be considered eligible to be assessed for either general skills or discipline-specific skills at the second point of testing. Achieving consensus is critical to maintaining comparability and credibility of VAA results across institutions/programmes. Because of this, it is likely that including a VAA component in the AHELO main study will require an extended testing window.

#### 4. Contextual survey

Overall results from the AHELO feasibility study and psychometric analyses of item level data should be used to refine the instruments that generate the contextual information. Design of the modified instruments should also take into account the kinds of predictors likely to be employed in the VAMs, as well as the institution/programme characteristics of policy interest.

#### 5. VAM selection

The Expert Group did not take a position on the selection of a particular VAM or VAA strategy. As indicated above, each model has advantages and disadvantages. Reflecting findings in the literature, general concerns were expressed concerning potential sources of bias including measurement error in the predictors and, more consequentially, (differential) selection bias and unobserved heterogeneity. The Expert Group agreed that interpretations should be made cautiously and with due regard to other relevant information, whatever the VAM employed.

#### 6. Comparisons using VAMs (National)

Presumably, a VAA would be conducted for a number of sets of institutions/programmes in a jurisdiction. Institutional and programme characteristics may be incorporated directly in the model or introduced in a second phase analysis. The results could then be used to generate hypotheses about the characteristics of more or less effective institutions or programmes. If institutions/programmes differ considerably on key contextual characteristics (e.g. size, selectivity, attrition rates, etc.), then, provided there are sufficient numbers of units to be compared, these characteristics can be used to form strata with comparisons conducted within strata. In any case, comparisons for benchmarking should always rely on multiple sources of evidence and not on a single quantitative indicator.

#### 7. Comparisons using VAMs (International)

The Expert Group recognized that different jurisdictions might well prefer different VAMs for use by their institutions. That prospect raises a number of questions: Who would do the work and carry out the requisite quality assurance? How would the results be reported? In that case should the OECD carryout a VAA for each institution using a common model within a jurisdiction and, if so, which model and how would the results be reported and used? What would be the implications for the data collection of creating common measures across all countries? Under what circumstances would comparisons across national boundaries be possible and desirable? What kinds of comparisons should be cautioned against or avoided entirely? Since these questions involve political and logistical, as well as technical issues, they could not be dealt with satisfactorily by the Expert Group. It was pointed out, however, that if one of the goals of AHELO was to facilitate international benchmarking, then such use of VAA results would be problematic for a number of reasons, including: *i)* differences in the



instruments used to measure prior ability; *ii*) likely variation among jurisdictions in the interpretation of the contextual surveys questions and *iii*) systematic differences in the amount of criterion variance accounted for by the predictors in the regression models.

#### 8. Student Samples

In the reports on the AHELO feasibility study, it has already been noted that obtaining representative samples of students from each institution and using every means to induce students to exert maximum effort is key to obtaining useful results at the institutional level. This is certainly true for VAA since differential motivation/effort contaminates estimates of both achievement and progress. For VAA, the problem of student attrition is of equal concern. In some institutions/programmes attrition rates can be considerable so that the students tested at the later time point constitute a highly non-random sample of the starting cohort. In that case, as indicated above, the usual interpretation of the value-added estimates must be modified and may not be regarded as particularly useful.

#### Summary

The discussions of the Expert Group ranged over a broad set of issues related to VAA. The general consensus was that, in principle, the inclusion of VAA in a future AHELO would be welcome, provided that a number of technical and logistical challenges could be overcome. Those challenges are not inconsiderable and so the decision to include VAA would have to be made at the outset as it would influence a number of design and cost factors for AHELO. At the same time, the Expert Group noted the difficulties inherent in making comparisons across jurisdictions, so that the main use of the results of VAA would be for institutional self-study and between programme comparisons within a jurisdiction. Whatever approach to VAA might be taken, emphasizing the cautions on interpretation and use would be essential.

**Meeting Participants**

Henry I. Braun, Chair	<a href="#">Boston College</a>
Daniel F. McCaffrey	<a href="#">Educational Testing Service</a>
Jeffrey T. Steedle	<a href="#">Council for Aid to Education (CAE)</a>
Julián P. Mariño	<a href="#">Instituto Colombiano para la Evaluación de la Educación</a> ICFES
Ou Lydia Liu	<a href="#">Educational Testing Service</a>
Peter T. Ewell	<a href="#">National Center for Higher Education Management Systems (NCHEMS)</a>
Richard Arum	<a href="#">New York University</a>
Timothy Rodgers	<a href="#">Coventry University</a>
Torbjørn Hægeland	<a href="#">Statistics Norway</a>
Diane Lalancette	<a href="#">OECD</a>

### Meeting Agenda

#### AHELO Experts meeting on value-added measurement

held at OECD Washington Centre,  
2001 L Street, N.W., Suite 650, Washington D.C. 20036-4922

from 9:00 to 17:00 on 31 January and 1 February 2013

#### Thursday 31 January (Morning session), 9:00 – 12:30

- 9:00 – 09:30**      **Welcome introductions**
- 09:30 – 10:30**      **Overview of the AHELO feasibility study and background context for the reflection on value-added measurement**
- **“Presentation / Questions and answers**
- 10:30 – 11:00*      *Coffee break*
- 11:00 – 12:30**      **Value-added measurement experiences in K-12 education**
- **Presentation / Questions and answers / Experience sharing**
- 12:30 – 13:30*      *Lunch*

#### Thursday 31 January (Afternoon session), 13:30 – 17:00

- 13:30 – 15:00**      **Value-added measurement experiences in higher education**
- **Presentation / Questions and answers / Experience sharing**
- 15:00 – 15:30*      *Coffee break*
- 15:30 – 17:00**      **Discussion**
- **Strengths and weaknesses of various value-added models used in education**
  - **Technical and political issues in the implementation of value-added measurement and the use of its results**
  - **Considerations in using value-added measurement in higher education**

Friday 1 February (Morning session), 9:00 – 12:30

- 9:00 – 10:30**      **Scope for developing value-added models in the context of an eventual AHELO main study**
- **Should an AHELO main study include value-added models?**
  - **Expected benefits and potential side effects of value-added models in an AHELO main study**
  - **Political issues in the use of value-added models and the use of their results**
- 10:30 – 11:00*      *Coffee break*
- 11:00 – 12:30**      **Experts' recommendations for an appropriate value-added model in the context of an eventual AHELO main study and methodological requirements**
- **Appropriate methodology for an AHELO main study**
  - **Assessment design (assessment cycle, sample design, data requirements)**
  - **Benefits and limitations of the proposed methodology**
- 12:30 – 13:30*      *Lunch*

Friday 1 February (Afternoon session), 13:30 – 17:00

- 13:30 – 15:00**      **Experts' recommendations for an appropriate value-added model in the context of an eventual AHELO study and methodological requirements - continued**
- 15:00 – 15:30*      *Coffee break*
- 15:30 – 16:30**      **Discussion about the group report**
- **Structure of the report**
  - **Process for completion of the report / Timeline**
- 16:30 – 17:00**      **Overall recommendations and wrap up of the meeting**

## REFERENCES

- Ballou, Sanders, and Wright (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, 37-65.
- Banta, T. W., & Pike, G. R. (2007). Revisiting the blind alley of value added. *Assessment Update*. 19(1), 1-2, 14-15. San Francisco: Wiley Periodicals Inc.
- Bird, S., Cox, D., Farewell, V., Goldstein, H., Holt, T., and Smith, A. (2005). Performance indicators: good, bad, ugly. *J. Royal Statistical Society A*, 168, Part 1, pp.1-27.
- Braun, H. (2005). Using student progress to evaluate teachers: A primer on value-added models. Princeton, NJ: Policy Information Center, Educational Testing Service.
- Braun, H., Kirsch, I. and Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12<sup>th</sup> grade NAEP reading assessment. *Teachers College Record*, 113, pp. 2309-2344.
- Council of Independent Colleges. (2008). Evidence of learning: Applying the collegiate learning assessment to improve teaching and learning in the liberal arts college experience. Washington, DC: Council of Independent Colleges.
- Domingue, B. (2012). Measuring effects of post-secondary institutions on student learning: A case study. (Unpublished manuscript).
- Espeland W. N., Sauder, M. (2007). Rankings and Reactivity: How Public Measures Recreate Social Worlds *American Journal of Sociology*, vol. 113, no. 1, pp. 1-40.
- Foley, B. and Goldstein, H. (2012). Measuring success: League tables in the public sector. London: The British Academy.
- Goldhaber, D., Walch, J., and Gabele, D. (forthcoming). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics, Politics and Policy*.
- Goldstein, H, and Speiglehalter, DJ. (1996). League tables and their limitations: Statistical issues in the comparison of institutional performance. *Journal of the Royal Statistical Society, Series A*, 159(part 3) 385-443.
- Haegeland, T, B. Bratsberg, L.J. Kirkeboen and O. Raaum (2011): Value-added indicators: A useful tool in the assessment of schools? Report 42/2011, Statistics Norway (In Norwegian).
- Hargreaves, A. and Harris, A. (2010). Performance beyond expectations. (Unpublished manuscript).

- Hess, F., & Finn, C. (2007). *No remedy left behind: Lessons from a half decade of NCLB*. Washington, DC: AEI Press.
- Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review*, 31(5), 415-439.
- Lachin, J.M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials*, 21, 167-189.
- Liu, O. L. (2008). *Measuring learning outcomes in higher education using the Measure of Academic Proficiency and Progress (MAPP™)* (ETS Research Report Series RR-08-047). Princeton, NJ: Educational Testing Service.
- Liu, O.L. (2011). Value-added assessment in higher education: A comparison of two methods. *Higher Education*, 61(4), 445–61.
- Liu, O. L., Bridgeman, B. & Adler, R. (2012). Learning outcomes assessment in higher education: Motivation matters. *Educational Researcher*, 41, 352 - 362.
- National Research Council (2010). Getting value out of value-added. H.Braun, N. Chudowsky and J. Koenig (Eds.) Washington, DC: National Academies Press.
- Organisation for Economic Co-operation and Development. (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. Paris: OECD.
- Organisation for Economic Co-operation and Development. (2012). AHELO: Feasibility Study Report (Vol.1:Design and Implementation). Paris: OECD.
- Organisation for Economic Co-operation and Development. (2013a). AHELO: Feasibility Study Report (Vol.2: Data Analysis and National Experiences). Paris: OECD.
- Organisation for Economic Co-operation and Development. (2013b). Literature Review on value-added. Paris: OECD
- Reyes, P., and Rincon, R. (2008). The texas experience with accountability and student learning assessment. In V. M. H. Borden & G. R. Pike (Eds.), *Assessing and accounting for student learning: Beyond the spellings commission: New directions for institutional research, assessment supplement 2007* (Vol. 2008, pp. 49-58). San Francisco, CA: Jossey-Bass.
- Rodgers, T. (2007). Measuring value-added in higher education: A proposed methodology for developing a performance indicator based on the economic value-added to graduates. *Education Economics*, 15(1), pp. 55-74.
- Saavedra, A. R., & Saavedra, J. E. (2011). Do colleges cultivate critical thinking, problem solving, writing and interpersonal skills? *Economics of Education Review*, 30(6), 1516-1526.
- Sauder and Espeland (2009). The Discipline of Rankings: Tight Coupling and Organizational Change. *American Sociological Review*, vol. 74, no. 1, pp. 63-82.
- Shavelson, R.J. 2009. *Measuring college learning responsibly: Accountability in a new era*. Stanford, CA: Stanford University Press.

- Shermis, M. D. (2008). The Collegiate Learning Assessment: A critical perspective. *Assessment Update*, 20(2), 10-12.
- Steedle, J. T., Kugelmass, H., & Nemeth, A. (2010). What do they measure? Comparing three learning outcomes assessments. *Change*, 42(4), 33-37.
- Steedle, J. T. (2012). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education*, 37(6), 637-652.
- Sutherland, P., Dueweke, A., Cunningham, K., & Grossman, B. (2007). Multiple drafts of a college's narrative. *Peer Review*, 9(2), 20-23.
- U.S. Department of Education. (2006). A test of leadership: Charting the future of us higher education. Washington, DC: U.S. Department of Education.
- U.S. News and World Report (2012). National University Rankings. Retrieved March 2, 2013 from <http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-universities>

## CHAPTER 11 - CONFERENCE REPORTING

The feasibility study was truly a learning experience and as such the conference was not so much there to present its results but rather to foster discussion on the questions raised. Because the feasibility study is discussed in details in volumes [1](#) and [2](#) of this Report, this Chapter will aim to limit redundant information and will try to concentrate on the issues discussed at the conference. The conclusions expressed aim to reflect the views presented by all conference participants.

The AHELO feasibility study Conference took place at OECD Headquarters in Paris on 11 and 12 March 2013. The conference was opened by Barbara Ischinger, Director for Education and Skills at the OECD and a keynote speech by Jamie P. Merisotis, President & CEO, Lumina Foundation for Education on “the emergence and rationale for measuring learning outcomes” (the full speech is available on the web at [http://www.luminafoundation.org/about\\_us/president/speeches/2013-03-11.html](http://www.luminafoundation.org/about_us/president/speeches/2013-03-11.html)).

A first day of plenary presentations and panels followed on:

- Lessons on what worked, what didn't work and what we learnt.
- What we learnt on about the purpose and uses for measures of learning outcomes.

This first day was followed by a second day of small-group workshops to discuss three main questions:

- How can international measures of learning outcomes provide a valid and valuable response to today's higher education challenges?
- What are the key challenges in developing an international measurement of learning outcomes?
- How can we combine an assessment that is useful to institutions with wider policy goals?

The agenda of the conference is provided in Annex E.



**Measuring learning outcomes: what for and for whom?**

*Highlights of the presentation by Andreas Schleicher, Deputy Director for Education and Skills, OECD*

There is a general agreement that more transparency about the performance of universities is a goal to strive for. But that is typically where the consensus ends. The role of student learning outcomes in measures of institutional performance is a complex one. Some consider that because education is ultimately about student learning that if we can measure student learning outcomes directly, we have got to the heart of institutional performance. Others point out that measuring something does not improve it.

There are important questions which need to be kept in mind: Why, exactly, do we need measures of student learning outcomes? And what, exactly, is it we are looking for? And how will we recognize it when we found it? This is the difficult question around issues of the validity of any such measures.

The search for measures of student learning outcomes is motivated by very different agendas, ranging from improving teaching through accountability up to demonstrating excellence to improve the standing of a university in a tough competitive market. Defining outcomes is complex, and has many dimensions.

How will we know that student test scores actually reflect the learning outcomes we are looking for? And how do we know that those learning outcomes are what will matter for the future success of students in life? This has to do with both construct and predictive validity.

**Why is information on student learning outcomes so important?** It has to do with the role information plays in today's societies. In some way information feeding peer pressure and public accountability has become more powerful than legislation and regulation.

There was a time when we would turn to universities to make judgements about the quality of universities. Today, it is the public that wants to make judgements about the quality of universities, and it clamours for reliable accountability tools to make those judgements. But there is also an important improvement agenda. It is difficult to improve what is not measured.

Comparing student learning outcomes can help individuals make better informed choices and employers to assess the value of qualifications, universities to understand their comparative strengths and weaknesses, and policy makers to quantify stocks and flows of high level skills and to assess value for money.

This is not easy. The biggest challenge is to define and operationalize higher education learning outcomes in ways that are valid across programmes, institutions, sub-systems and cultures. But you also pay a high price for not doing this. Without such data, judgements about higher education outcomes will continue to be made on the basis of idiosyncratic rankings derived from higher education inputs.

At the end of the day, student learning outcomes are the bottom line of today's universities. There is simply no longer a need to go to college to study, to get a degree, or to meet with a professor. And there is no longer a need to go abroad to study internationally.

Together, all this provides a robust motivation to pursue the agenda of measuring student learning outcomes.

But how, exactly, would we define these elusive student learning outcomes? **What** do we need to assess, and **whom** do we need to tell about the results?

AHELO struggled with some important design questions. Surely, policy makers are interested in learning something about their higher education system as a whole. But when looking at this carefully, we see that there is simply too much variation in institutional structures across countries. It is also rather unrealistic to obtain nationally representative samples. More important, even if you could mandate such assessments, it is unlikely to be effective as a tool for improvement at the level of service provision.

That is why AHELO set out to measure outcomes at the level of institutions, departments and faculty. The idea of AHELO is to combine the definition of an OECD measure of quality with reliable assessment methods to which institutions can voluntarily subscribe and which might progressively find wider acceptance. It is a pragmatic solution, but one that will not yield system-wide insights.

And then there is the question of **what you want to assess**. Some say that one should focus assessments on established disciplines, such as engineering or economics. Those are easily interpretable in the context of departments and faculties. However, it is not straightforward and requires highly differentiated instruments, and excludes competency areas that are not amenable to large-scale assessment or not sufficiently invariant across cultures and languages.

Others say that the focus should be on transversal skills. Surely, those are less dependent on occupational and cultural contexts, they should be applicable across universities, departments and faculties, and we know that they are powerful drivers for improving the quality of teaching in the disciplines too. But there are important drawbacks too. Those transversal skills reflect cumulative learning outcomes and therefore need to be related to prior learning. They also do not relate to the kind of subject-matter competencies that many universities, departments or faculties would consider their province.

Again, AHELO took a pragmatic approach and assessed a combination of both, which leaves the conceptual issues for later to resolve.

And then there is the issue of what kind of information do we need **for whom**. Individuals, whether prospective students or employers, would want to know the “bottom line” of the performance of institutions, departments or faculties, so they are interested in actual scores. By contrast, institutions and policy makers wishing to assess the quality of services provided would be interested in the “value added” by the institutions. Nobody has squared that circle yet. Last but not least, there is this complex relationship between evaluation and trust. In sum, establishing what kind of outcomes we need for whom is far from straightforward.

So how will we know that outcomes measures provide meaningful, **valid and robust results**?

What would we expect from such outcome measures? Surely, those measures should reflect central and enduring parts of higher education teaching that relate to quality of outcomes.

They should reflect aspects that can be improved, as there is no point measuring things you cannot do anything about. And, when we are talking about international assessments, we need to make sure that our measures are cross-culturally appropriate and valid across institutions and systems.

There is also the difficult balance between the **breadth and depth** of such measures. You want to make sure that you cover a sufficiently broad set of skills to avoid the kind of tunnel vision that leads to narrowing the curriculum. At the same time, you want to ensure that the individual components of your measures are sufficiently deep to provide meaningful feedback to students and faculty. You also want to make sure that your measures provide a powerful communication tool that can stimulate improvement. And you want to ensure that your measures are as comparable as possible but as specific as necessary to reflect the context in which they are interpreted.

These demands on the measures shape **what we need to expect from robust assessments**. First of all, these should support improvement of learning at all levels of the education system. Second, you want to go beyond seeing whether students get a question right or wrong, so your assessments should make students' thinking visible and allow for divergent thinking. We all know how rapidly the demand for skills in our societies and labour-markets change. Some people say that if you want to measure change, you cannot change your measure. But if you freeze your measure over time, your assessments will become stale very quickly. So it is important to ensure that assessments are adaptable and responsive to new developments. But it is also important that they add value for teaching and learning by providing information that can be acted on by students, teachers, and administrators. The assessments also need to yield interpretable scales and you want them to be largely performance based.

And let's not forget some tough **methodological challenges**. Can we drink from the fire hose of increasing data streams that arise from new assessment modes? Can we ensure that the essence of what we measure does not get lost in increasingly sophisticated contexts for tasks? How can we create assessments that are activators of students' own learning? Can we utilise new technologies to gain more information from students without overwhelming students with more assessments? And how will we balance crowd wisdom and traditional validity information when we assess the relevance of assessment tasks? And then the ultimate test of truth is whether our assessments line up with what we actually want to predict with them.

So where do we start to get all this done? The keywords here are **coherence, comprehensiveness and continuity**. Coherence means that we build on a well-structured conceptual base—an expected learning progression—as the foundation for assessments, and to ensure consistency and complementarity across administrative levels of the system. Comprehensiveness means that we need a range of assessment methods to ensure adequate measurement of intended constructs and measures of different grain size to serve different decision-making needs. We also need to think about producing productive feedback, at appropriate levels of detail, to fuel accountability and improvement decisions at multiple levels. And continuity means to provide a continuous stream of evidence that tracks progress.

AHELO has started to pursue this agenda, but we also see that we are still at the beginning of a long path.

**So what does all of this mean?** Surely, student learning outcomes must be in the critical path of assessing the outcomes of higher education. AHELO has shown that we can test some of these internationally. What we don't yet know is what part of the bigger picture on student learning outcomes tests and assessments of this kind can and should be.

### The feasibility study: a starting point

The experience of the AHELO feasibility study and the main lessons learnt from it have been developed in detail in volumes [1](#) and [2](#) of the feasibility study report. This experience was the basis for the presentations and discussions during the conference and fed into the wider debate around an international assessment of learning outcomes. It is important to keep in mind though that the purpose of the feasibility study was to provide a proof of concept: was it technically and practically feasible to assess what students know and can do near graduation? Because of the limitations inherent to a feasibility study, the analysis of the results only pertain to the instruments tested and cannot be generalised.

This said, we can conclude that the feasibility study has been successful in providing this proof of concept but also in bringing to the forefront the issue of learning outcomes. The questions brought up by the process of the feasibility study are as much a part of its success as the actual findings and were the focus of the conference.

« AHELO is a starting point. We started the discussion. It started with quality assurance and quality improvement is the next step. »

*Michael Hoffmann, SEFI (European Society for Engineering Education)*

### What purposes are attributed to AHELO?

**Exercise 7** of the workshops asked participants about the uses of an evaluation of learning outcomes. The answers from participants ranged from improvements to teaching and the curriculum, identifying best practices, comparisons and benchmarks. The detail of these answers is provided in Annex F.

Throughout the conference participants also expressed their views on what an AHELO should be and what purposes it should fulfil. Participants thought that an AHELO could:

- Be a tool for benchmarking quality.
- Be a tool that yields qualitative data.
- Be a tool that might better align graduate outcomes to labour market needs.

- Be a tool for accountability.
- Be a measure of how well a programme is progressing.
- Be a potential source of information for programme improvement (if the correct data are made available).
- Contribute to preparing a qualified international workforce to meet global demands and societal needs.
- Help governments to better direct their resources.

One of the lessons of the feasibility study was that it is essential to clearly communicate on the purpose of an AHELO. It is important therefore to remember that an AHELO as envisaged in the context of the feasibility study was **not**:

- A measure of HEIs overall performance.
- A measure of country higher education performance.
- Designed for accountability.
- Designed for ranking.
- A measure of teacher or teaching quality.
- A measure for assessing individual students.
- A qualification for students.

Rather, AHELO was envisaged as a low-stakes exercise geared at institutions to inform diagnosis and improve their teaching in light of this evidence.

There is also a need to be very specific on the formative value and the specific feedback that institutions and countries would receive, with clear terms of engagement and deliverables, outcomes and services provided.

The stated aim of AHELO was to provide higher education institutions with feedback on the learning outcomes of their students, which they could use to foster improvement in student learning outcomes. Some countries want AHELO to serve a public policy goal beyond providing insights for the institutions assessed. How these wider goals could be addressed without compromising the low-stakes, formative nature of AHELO would need to be considered.

#### **An evaluation within the wider environment of higher education**



*No single aspect of life has more world-changing potential than education.*  
*Jamie Merisotis, President & CEO, Lumina Foundation for Education*



As expressed in the first two Chapters of the Report (Volume 1) the need for an AHELO has developed within an evolving higher education context. There are many challenges facing higher education today: doing more (and better) with less; what the institutions are expected

to focus on and achieve; a growing and more diverse student body; globalisation and internationalisation, etc. (see **exercise 1** in Annex F to see in more details what conference participants identified as the main challenges facing higher education today).

An evaluation like AHELO needs to be well integrated in the wider policy dialogue and we need to fully understand and work through the implications of higher education complexities before moving on to the technical development of instruments.

Within this context a question that comes up repeatedly is the issue of quality. It is within this quality challenge that the necessity of a direct evaluation of student learning outcomes has developed. Maintaining and improving quality in teaching and learning is important to HEIs, students, employers, and governments.

« *Learning outcomes is the core of what quality assurance bodies care about. Nothing is more important than what students can do and having a reliable measurement of those outcomes which have been achieved.* »  
*Harvey Weingarten, HEQCO (Higher Education Quality Council of Ontario)*

Participants agreed that an AHELO would need to be one component of a multi-layered system of assessment. It could not possibly look at everything higher education strives to offer or have all the answers. AHELO is also in line with a number of other initiatives being developed with the concept of learning outcomes measurement in mind (Tuning, Quality Frameworks, etc.).

« *This may be one part of the puzzle, but is not enough.* »  
*David Robinson, Education International*

### **What role should stakeholders have?**

An important lesson of the feasibility study (see Chapter 2) and one that was stressed repeatedly by conference participants is the necessity of a collaborative aspect to the study. All stakeholders need to be as involved as possible and as early as possible.

« *Let's consider students not as consumers but as equal partners who have their own say.* »  
*Nevena Vuksanovic, ESU (European Student Union).*

Participants emphasized that there would be a great benefit if participating countries, institutions, faculty, students and business were able to contribute to the design of the study beyond the level of what could be achieved within the feasibility study, as their perspective is invaluable. Also implicating faculty and students was one of the solutions proposed to increase response rates.

### Is there an added value for an international assessment?

Numerous initiatives are taking place at the national level. For many of the countries who took part in the feasibility study this work was but an extension of the evaluation work done at the national level. But today's economy and the greater mobility of students, faculty and workers have increased the interest in the "performance" of institutions on a global scale.

Existing international tools are proxy measures of learning outcomes (e.g. rankings). A tool like AHELO could allow institutions to benchmark their performance against other institutions not just locally but internationally.

#### ***Learning from each other: the experience of the feasibility study participants***

« As we've delved into this area over the last few years, my Lumina colleagues and I have remarked again and again on the intercontinental "ping-pong" effect of this work. What is learned in Melbourne or Leiden or Shanghai informs our efforts in Boston, Miami and San Francisco [...] and the steps that we take then seem to alter the paths taken in those places as well. In short, we all learn as we do this work, and we adapt and use those lessons in our own particular contexts. »

*Jamie Merisotis, President & CEO, Lumina Foundation for Education*

The countries who participated in the feasibility study gave their views on the experience in the second Volume of the Report. The diversity of participants added challenge but also richness to the feasibility study. One of the true benefits of the feasibility study has been to get countries together to discuss and compare the way they do things.

Although some frustration was still present at the time of the conference because countries had not yet received all the data they needed to proceed with further analysis (and because the institutional reports prepared by the Consortium were not up to the high expectations of institutions and would need to reflect their strengths and weaknesses in the educational environment to help with improvement), overall the view of participating countries was that AHELO had been a useful experience and that much was learnt along the way, at the international level but also at the national level.

A few countries noted changes to the curriculum which were already taking place as a result of their participation. For example some faculty in Japan and Ontario stated that the constructed response tasks of the AHELO feasibility study helped them reconsider how they teach. It has generated reflection about programme content, curriculum, delivery, and assessment. Another concrete example in Italy: one of the HEIs who participated in the feasibility study has since decided to start introducing econometrics in the first three years of the economics degree. The main quality assurance agency also decided to introduce a generic skills test for 20 000 students (with problem solving, comprehension and critical thinking).

Participating in the feasibility study has also had the additional advantage of building technical capacity in administering large-scale assessments. The countries were all at different levels of

preparation (and some were working with very short timelines) but all managed to put in place the structures necessary to a successful implementation. One issue to note is that an international assessment brings with it the built-in difficulty of translating and adapting the test instruments to the languages and cultures of the participating countries. This process is detailed in Volume 1 of the report.

### **What's in it for us? How to motivate participants?**

For an assessment like AHELO to work, engagement and motivation are needed at all levels: from policy makers to institutions to individual faculty and students. For all involved participation in such an endeavour means an investment of time and energy. What each participant can get from their participation needs to be clearly identified and communicated. Even with the best test in the world, results will be meaningless if it is not possible to get students to sit down and take it (and give it their best effort).

**Exercise 8** of the workshops was a role-playing exercise where participants were asked to pretend they were a student, an academic dean or head of department, the person responsible for international affairs at an institution, a policy official in a higher education ministry, an employer, a faculty member, a university president, a rector or vice-rector, or a higher education researcher. For each of these categories the participants had to come up with key points on the interest of an AHELO. Please see Annex F for the suggestions from workshop participants.

Through the course of the feasibility study the level of interest for AHELO at the level of **policy makers** was notable. There was a high level of enthusiasm and commitment for AHELO in the participating countries. One notable example is Egypt which managed to successfully implement the assessment in all three strands in the midst of a revolution.

**Institutions** were also quite receptive to participating in AHELO. Both in Italy and Ontario, for example, almost all the institutions approached were very interested in participation. Institutions want to see how they are performing compared to sister institutions elsewhere. The future of an AHELO will in great part depend on what institutions feel it can bring to them.

Clearly communicating on AHELO and its purposes is important in getting **faculty** involved. Taking faculty time away from teaching and research requires a very good reason, as well as a clear description of how the activity might benefit institutions. But faculty involved in the feasibility study clearly saw its potential for teaching and learning.

One of the biggest challenges of the feasibility study for some countries was motivating a sufficiently high number of **students** to take the test. Since AHELO is not a high-stakes exam, students might not quite see why they have to take time to do it when they are busy otherwise.








### Motivating students in Italy (by Fiorella Kostoris)

Three incentives seemed to work in Italy not only to raise the response rate, but also to induce the best students' efforts once their decision to participate was made (and we know that the latter are highly correlated with good results).

1. To motivate their desire to get a self-evaluation: you know your test results after scoring and you can compare yourself with various benchmarks.
2. To motivate their desire to get a certification: you know your test results after scoring and you may ask for a certificate useful for labour market purposes.
3. To motivate their desire to provide an assessment of their University: you know your Department's results after scoring and this gives you a benchmark and an element of comparison between the quality of your studies and institution relative to that of others elsewhere.

We asked conference participants to tackle this issue in **exercise 6** of the workshops. Workshop participants strongly suggested giving students their results (even going one step further and finding a way to use the responses as a learning tool in the classroom) or to give them some kind of credit for their participation.

Please see below the top five answers (the figure is the number of times this was mentioned by participants):

-  **36** Feedback to the students of their results (including discussing their mistakes with faculty)
-  **34** Give a credit or certificate
-  **22** Embed testing in the curriculum/existing exam
-  **21** Clear and detailed explanation of the project and its importance within the global movement for evaluating learning outcomes
-  **19** Monetary incentive (for example reduced module fees)

For the rest of the list of suggestions received, please see Annex F.

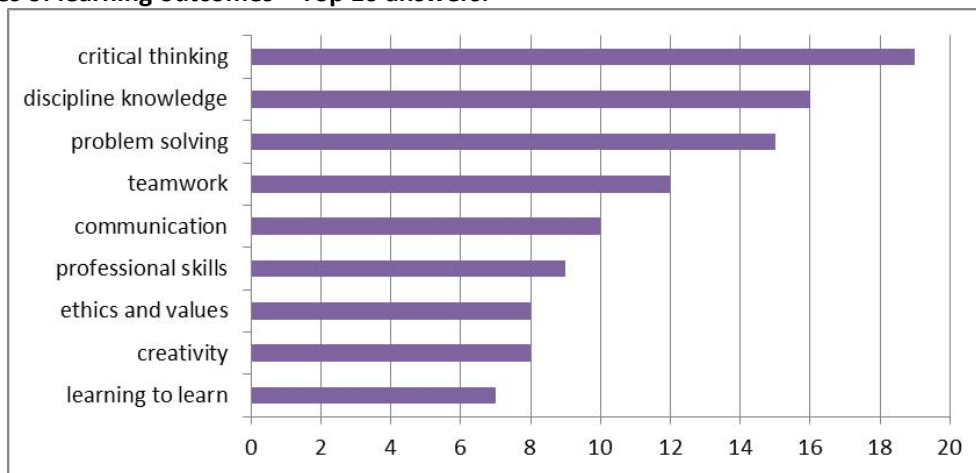
## What do you assess?

### *Types of learning outcomes*

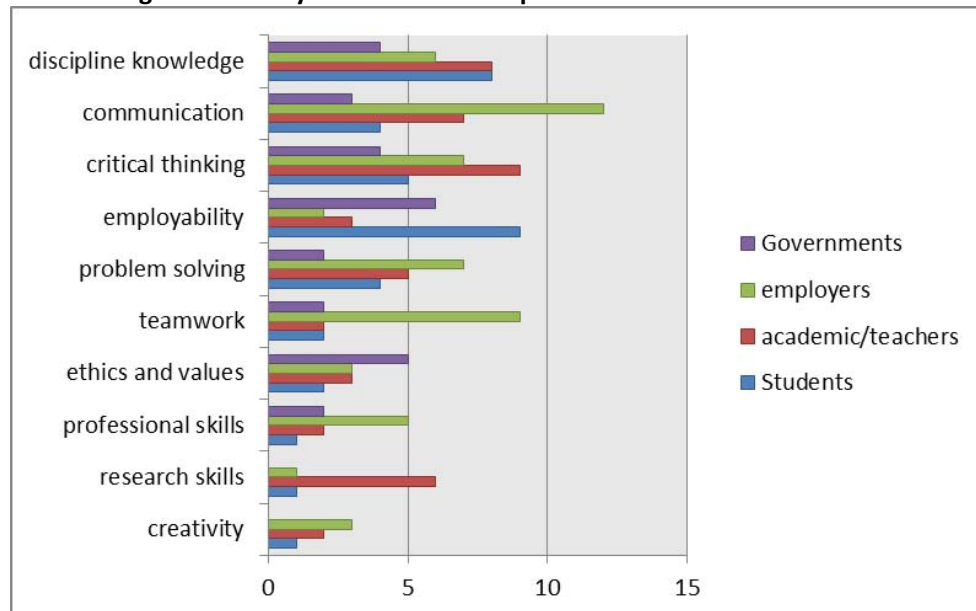
The first step in developing an AHELO is to define exactly what it is we are trying to measure.

In **exercise 2** of the workshops participants were asked to identify the most important learning outcomes of higher education. Most groups came up with quite a few suggestions. Almost all had a mix of discipline skills and generic skills such as problem solving, teamwork, or communication.

#### Types of learning outcomes – Top 10 answers:



#### Types of learning outcomes by stakeholders – Top 10 answers:



See Annex F for the complete list.

The question of the relevance of skills to the labour market was also raised during the conference. Should an assessment focus on these learning outcomes that are valued by employers? Would this be perceived negatively by institutions and faculty?

#### *Generic skills or discipline-specific skills*

The choice of assessment strands is complex. For the sake of the feasibility study, generic skills and discipline specific skills were assessed separately. This was a way to test different approaches. For a future assessment this separation may not be the method chosen.

**Exercise 4** of the workshops asked the groups their opinion on the strengths and weaknesses of a generic skills, discipline specific or blended approach, and suggestions on how to achieve this blended approach. The complete answers are provided in Annex F and summarised below.

Testing **generic skills** was generally considered to be a challenge at the test development level. However participants felt that it would apply to different populations (with some limitations). The results though complex and limited would have great potential for analysis and use. Generic skills were considered essential skills to have but should not be considered the be-all and end-all.

Testing of **discipline specific skills** was considered useful on a global scale but it was felt that the diversity of local contexts and disciplines would create difficulties. In general this type of testing was thought to be easier and cheaper if you test one discipline but the costs would add up for each discipline you add to the test. Achieving consensus could be hard work but the test could prove more intrinsically interesting and engaging for the participants, provided there is no oversimplification of the test (and the results remain relevant).

While several suggestions were put forward on the best way to achieve a **blended approach** the most prevalent answer was to find a way to assess generic skills within a discipline context. The blended approach was thought to provide better feedback, to be more useful and comprehensive as well as essential (evaluation of both generic and discipline skills going hand in hand). However participants also noted the added complexity and increased strain on time and resources which would result from this approach, while also expressing the concern that the results may be somewhat limited.

#### ***The contextual dimension***

The feasibility study emphasized the importance of a well-developed contextual dimension. When we asked the workshop participants in **exercise 7** which data would be most useful to HEIs the contextual data was very present. The importance of contextual data to help get the most out of the results of the assessment was also highlighted in the interventions of the experts and country representatives who participated in the feasibility study.

### ***Measuring value added***

« *The essence of universities/HEIs is what lasts forever for the graduates. Education is what is left after school, what the students take with them. How much do we transform the students? This is a challenge to assess.* »  
*Alfredo Dajer Abimerhi, Universidad Autónoma de Yucatán (UADY, Mexico)*

Again and again in the panel discussions and workshops the importance of this question was underlined. Although value-added measurement was not included within the feasibility study, the group of experts convened by the OECD has tried to give the first set of answers to the question of how we go about measuring this (see Chapter 10 for their report).

### **How do you measure learning outcomes?**

Once agreement has been reached on what it is exactly that we should measure, the next step to operationalize an assessment is to look at the ways of getting these measures both through the development of assessment instruments and through an efficient implementation of these instruments.

### ***Types of measures***

The feasibility study measured learning outcomes through a direct assessment of students. This is but one way to do this. One could also consider such measures as completion rates, national qualifications frameworks, quality assurance, or graduate employment outcomes for example.

With this in mind workshop participants were asked in **exercise 3** to suggest up to three measures of learning outcomes, as well as their strengths and drawbacks. Some participants did not feel that some of the proposals presented as examples were really measures of learning outcomes. But another way to look at this issue is that if employability is a learning outcome then graduate employment must be a measure. The different measures and their place will need to be considered.

The measures which were most often cited were:

- Surveys
- Labour market outcomes
- Student testing
- Quality assurance and accreditation
- Benchmarking and comparison

The many answers to the question (including strengths and drawbacks) are available in full in Annex F.

**CRTs or MCQs**

Within the feasibility study a mix of Multiple Choice Questions (MCQs) and Constructed Response Tasks (CRTs)<sup>1</sup> were used.

**Exercise 5** tackled the question of the strengths and drawbacks of each item type.

Overall the groups felt that MCQs:

- were easier to develop and administer;
- were more cost effective, more reliable and faster;
- allowed for more objective and easier scoring; and
- made comparisons easier.

On the downside they thought that MCQs:

- assessed a lower level of skills;
- raised questions about validity and development; and
- produced a test and results with limited interest.

Almost as a mirror image of this the strengths of CRTs were felt to be:

- the higher level of skills assessed;
- the interest of the test and results; and
- a more comprehensive test.

The drawbacks of CRTs were noted to be:

- more subjective scoring;
- issues on validity;
- a complexity of development and administration;
- increased costs; and
- the difficulties linked to culture and languages.

There are pluses and minuses in both approaches and there is no clear cut answer on whether one is better than the other. The balance of trade-offs needs to be considered. Some groups also put forward the suggestion of the portfolio approach.

On the specific issue of the CRTs the report from the TAG was that the experts support the inclusion of this type of task (with the caveat that the ones developed for the feasibility study proved too difficult). CRTs limit generalisability but they were interesting for participants. The contextualisation of these particular tasks was difficult. Multiple languages and contexts added difficulty and involved trade-offs to modify for context without changing the task too much. Students liked the CRTs, even if for some this was a new experience.

---

<sup>1</sup> Annex B of volume 1 of the feasibility study report includes illustrative items from the test.

### ***Timelines, test administration and delivery***

Important lessons were also learnt through the feasibility study on the implementation of such a test. Some of these points were raised by various participants in the interventions and discussions during the conference.

First and foremost was the issue of time. Participants noted that enough time and resources have to be devoted from the beginning to the development and implementation of the assessment. In the case of the feasibility study the prolonged planning period made it difficult to keep the higher education community interested and engaged. Enough time also has to be allotted for scoring as human scoring requires tools, training and careful monitoring to strengthen reliability. The timing of the testing is also important as it strongly impacts student response rates in a lot of cases. Therefore enough time has to be set aside to have the testing windows fit optimally in the academic year of the institutions participating.

Another important reminder was that reaching international consensus on the assessment framework is essential before the instrument development takes place.

The administration of the test was complex, on a large scale and all computer based. This, in the great majority, worked well and all participants concluded that it is indeed possible to deliver a test electronically to a large number of students within a reasonable time frame.

### **Was AHELO worth the cost and should it continue?**

Different participants had different opinions on this. A general feeling was that maybe more data and analysis was still to be gained from the feasibility study before such a judgement could be made. While the international costs have been clear and documented, it is not possible to get figures for the national costs due to the nature of the feasibility study. Before a main study could be envisaged a clearer idea of the full costs of such an assessment would be needed.

While the participants had many suggestions on the way institutions could use AHELO data this would still need to be further developed and clarified. The need for AHELO-type data and analysis is very much there however (see answers to **exercise 7** in Annex F).

« *This is not going away. The interest is growing. It is critically important that it is well done and exercises like this will help shape the solutions.* »  
*Harvey Weingarten, HEQCO (Higher Education Quality Council of Ontario)*

« *Business and industry had high expectations which were pretty much achieved. This is a tough exercise and there are excellent people working on the project. How to find the appropriate funding and how to prevent constant financial uncertainties will be important questions. We hope the work continues.* »  
*Marita Aho, Confederation of Finnish Industries, BIAC*



*Is AHELO a panacea? No. But it is contributing to the establishment of a shared vocabulary on quality and enhancing our work on continuous quality improvement. It has also provided some valuable lessons about the role learning outcomes can play in enhancing transparency, measuring quality and demonstrating the achievements of publicly funded institutions.*



*Deborah Newman, Deputy Minister (Canada, Ontario)*

## ANNEX E: CONFERENCE AGENDA

**Measuring learning outcomes in Higher Education:**  
*Lessons learnt from the AHELO Feasibility Study and next steps.*

OECD  
BETTER POLICIES FOR BETTER LIVES

OECD Conference centre, Paris  
11-12 March 2013

*Programme*

**ahelo** Assessment of Higher Education Learning Outcomes



**MONDAY 11 MARCH 2013****Plenary 1 – Conference opening****Welcome speech**

Barbara Ischinger, OECD Director for Education and Skills

**Opening keynote - The emergence and rationale for measuring learning outcomes**

Jamie Merisotis, Lumina Foundation for Education

**The making of the AHELO feasibility study and key findings**

Deborah Roseveare, Directorate for Education and Skills, OECD

Peter Ewell, Chair AHELO Technical Advisory Group

**Plenary 2 – Lessons on what worked, what didn't work and what we learnt from the Feasibility Study experience**

Chair/moderator: Peter Coaldrake, Vice-Chancellor, Queensland University of Technology (Australia), Chair of the IMHE GB

**Introduction and brief presentations on what worked and what did not**

Jan Levy, Chair of AHELO Group of National Experts

Satoko Fukahori, Japan

Saana Radi, Egypt

Fiorella Kostoris, Italy

**Panel discussion**

Peter Ewell, Chair of the TAG

Jan Levy, Chair of the AHELO Group of National Experts

Satoko Fukahori, Japan

Saana Radi, Egypt

Fiorella Kostoris, Italy

Diane Lalancette, AHELO team, OECD

**Plenary 3 – What we learnt about the purpose and uses for measures of learning outcomes?**

Chair/Moderator: Steve Egan, Deputy Chief Executive and Director (Finance and Corporate Resources), HEFCE (United Kingdom), Vice-Chair of the IMHE GB

**Keynote 2 - Measuring learning outcomes: what for and for whom?**

Andreas Schleicher, OECD

**Stakeholders' views on measuring learning outcomes****First Panel**

Harvey Weingarten, Higher Education Quality Council Ontario (HEQCO)  
Marita Aho, Confederation of Finnish Industries, BIAC  
Michael Hoffmann, SEFI (European Society for Engineering Education)  
Nevena Vuksanović, European Students Union  
David Robinson, Education International

**Second Panel**

Eva Egron Polak, Secretary General, International Association of Universities  
Roman Nedela, Matej Bel University (Slovak Republic)  
Kukio Kishimoto, Tokyo Institute of Technology (Japan)  
Alfredo Dajer Abimerhi, Universidad Autónoma de Yucatán (UADY, Mexico)  
Deborah Newman, Deputy Minister (Canada, Ontario)

**Plenary 4 – Taking AHELO forward: next steps and the importance of the workshop discussions**

Deborah Roseveare

<b>TUESDAY 12 MARCH 2013</b>	
<b>Workshop 1</b>	
	<b>How can international measures of learning outcomes provide a valid and valuable response to today's higher education challenges?</b>
<b>Workshop 2</b>	
	<b>What are the key challenges in developing an international measurement of learning outcomes?</b>
<b>Workshop 3</b>	
	<b>How can we combine an assessment that is useful to institutions with wider policy goals?</b>
<b>Plenary 5 – Conference closing</b>	

## ANNEX F: WORKSHOP EXERCISES

On the second day of the conference participants were split into groups of 10 and were given a series of questions to discuss. The feedback below aims to be synthetic while being as complete as possible by transcribing the answers as given by the participants.

### **Workshop 1 - How can measures of learning outcomes provide a valid and valuable response to today's higher education challenges?**

#### ***Exercise 1: Identify Challenges***

In the first exercise we asked the groups to identify challenges facing higher education today and to discuss how these challenges were affecting higher education. For better reading we have grouped the answers in six large groups.

<b>Assessing Learning Outcomes as part of Quality assessment</b>	
<b>Definition and assessment of Learning Outcomes</b>	affects teaching, learning and assessment processes as well as faculty development
<b>Quality</b>	provide information for QA systems on Learning Outcomes, benchmarking information
<b>LO culture</b>	
<b>Quality</b>	Affects family expectations, government allocations of funding, job opportunities and employability.
<b>Concept of education</b>	Do learning outcomes capture the purpose of higher education. Higher level concepts are often fuzzy.
<b>Quality assurance in HE (including assessment of HE outcomes)</b>	Resources adequacy for educational services/activities/processes. Fulfilment of quality standards requirements. Many tools to assess LO. Defining the prioritizing learning outcomes (reputational learning outcomes / industrial targeted learning outcomes, academic targets)
<b>Lack of self-assessment and comparability (benchmarking) tools for jurisdictions to know where they stand vis-à-vis their peers</b>	Affects transparency, strategic positioning and institutional development.
<b>Depth of indicators focused on education (beyond research)</b>	Research quality hijacks understanding of education quality.

<b>Massification, equity and quality</b>	
<b>Massification of HE</b>	Transferring the university system from elite education to mass education. The massification of HE causes "quality" issues and capacity of staff and university capacity also lack of funding to maintain quality
<b>Massification of HE, mobility</b>	Different student backgrounds. We need a mechanism that works in the marketplace, ensures quality as more and more people enter HE. And move HEIs and back and forth with working.
<b>Access</b>	Difficult to balance access and quality
<b>increased diversity in student bodies</b>	Learning Outcomes must be suitable for range of student abilities
<b>diversity</b>	affects LO attainment or achievement, curriculum design, teaching and learning as well as assessment methods
<b>expanding access</b>	Maintain quality. Learning Outcomes are only one element of quality assurance
<b>massification and quality</b>	learning outcomes must be most attentive to question of quality
<b>Demographic diversity of students</b>	Resources adequacy for educational services/activities/processes. Fulfilment of quality standards requirements. Many tools to assess LO. Defining the prioritizing learning outcomes (reputational learning outcomes / industrial targeted learning outcomes, academic targets)
<b>Mass higher education - impacts on quality especially standards eg 70-80% participation. Rapid expansion makes problem worse</b>	badly- standards are going down (both students and faculty are poorly prepared)
<b>Demonstrating quality (used to be assumed)</b>	Institutional autonomy challenge. Diverse student population with different learning outcomes. Agreeing a "neutral currency" which demonstrates student abilities across different institutions. Will employers/professional bodies change their recruitment practices? Getting the whole range of institutions to participate.
<b>Diverse student body</b>	Pay a lot more attention to the social dimension of HE and inclusion.

<b>Globalisation</b>	
<b>student mobility</b>	transferability of qualifications
<b>Internationalisation/ Globalisation</b>	causes emerging problems for degree equivalence and credit transfer, problem of quality assurance

<b>Adapting to change</b>	
<b>IT impact on teaching</b>	impact on teaching and quality, deepening learning, MOOCs
<b>Diversity/custom design</b>	Multiple providers for one degree

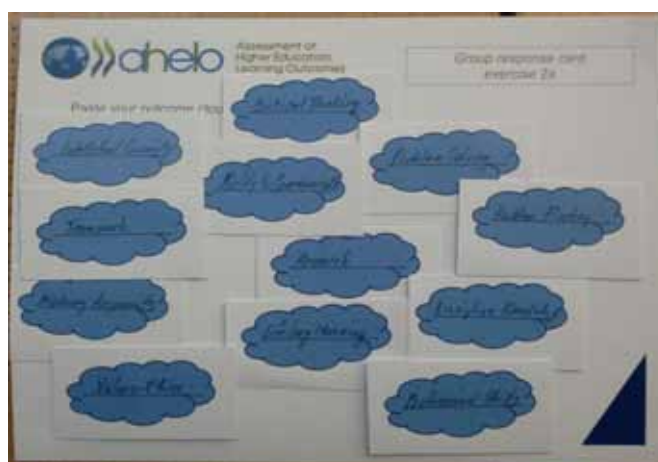
<b>Evolving and sometimes diverging expectations</b>	
<b>Readiness for labour market</b>	Student readiness for labour market e.g. system has problems motivating students for science and engineering
<b>Multiple roles of HEIs/Competing expectations</b>	Focus on employability to the exclusion of other LO
<b>Increased student expectations</b>	Due to high costs. Pressure to deliver high quality
<b>Sustainability of mission</b>	autonomy erosion, beyond financial, rapid growth affects quality
<b>Defining and improving quality</b>	assuring and/or improving
<b>Diverging expectations</b>	reaching agreement
<b>Student expectations</b>	outcomes continuously change
<b>Society outcomes and values</b>	The test needs to cover important areas which change by country and culture at the HEI and Student area.
<b>Political/Ideological implications</b>	can influence the design/expected outcomes
<b>Student expectations</b>	
<b>Employability of graduates</b>	Demonstrating and assessing the skills and knowledge employers want.
<b>Purpose of HE</b>	Funding influences purpose of higher education. Short term pressure on institutions influence HE. Standardisation? Do we want it or not?
<b>Aligning learning outcomes with rapidly changing job market needs</b>	Affects student preparedness for workforce, adaptability to changing skills.

Financing and the concept of accountability	
<b>financial duress</b>	affects quality
<b>Funding</b>	Quality and LO assessment: accountability vs accreditation. Possible funding distribution. Formative
<b>Financing HE, cost effectiveness in HE</b>	Resources adequacy for educational services/activities/processes. Fulfilment of quality standards requirements. Many tools to assess LO. Defining the prioritizing learning outcomes (reputational learning outcomes / industrial targeted learning outcomes, academic targets)
<b>Financing HE - links to accountability</b>	AHELO is expensive, i.e. international studies expensive. But accountability does not require an international study.
<b>Funding</b>	If there is more money we can deliver more quality knowledge. If the funding is decreasing then targets are decreasing as well. Demand for more money from government to HE in general.

**Exercise 2: What are most important learning outcomes in higher education?**

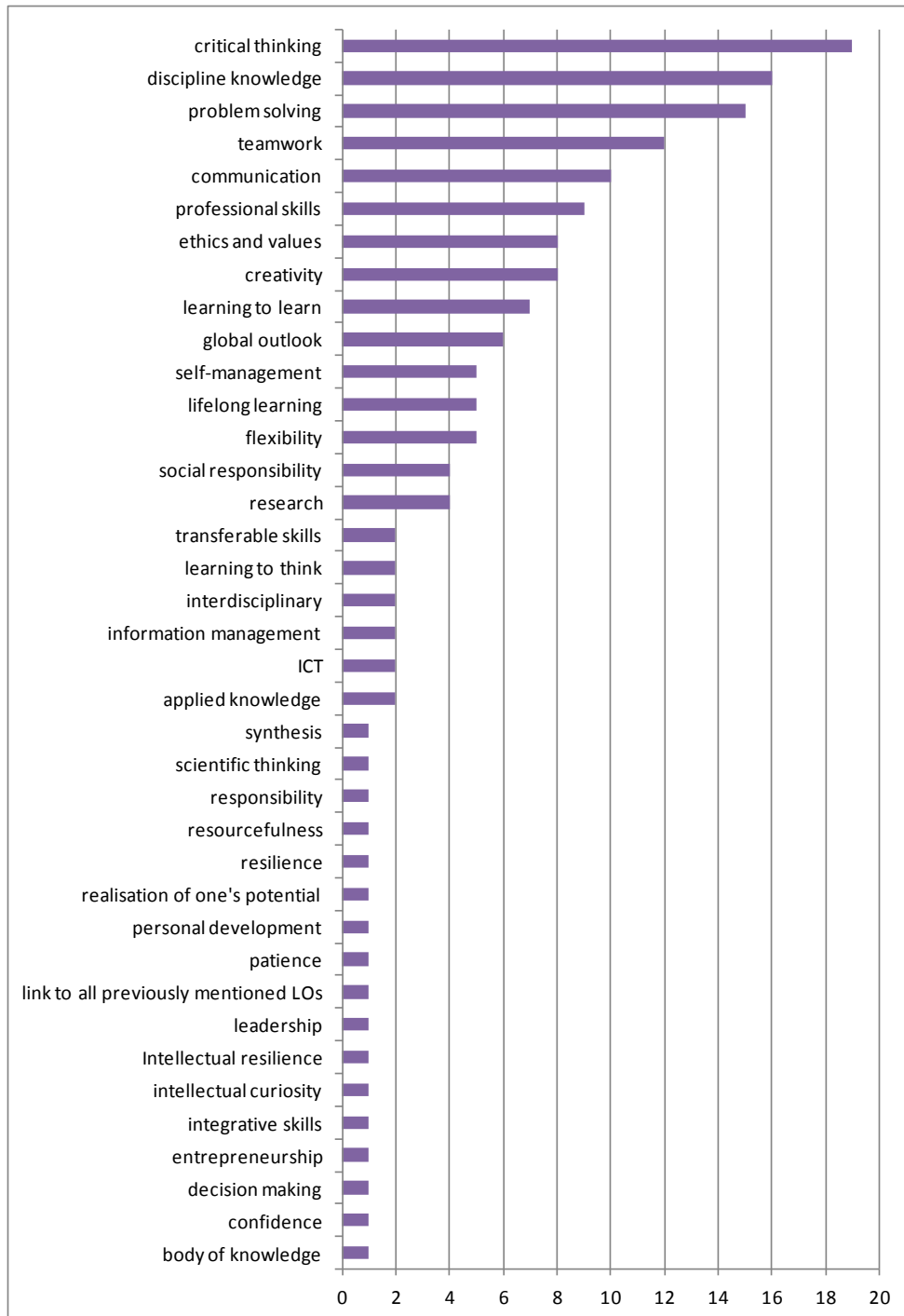
*Exercise 2a: Identify the major learning outcomes*

What do we mean by learning outcomes? To answer this question the groups were asked to identify a list of learning outcomes and write those out on their response cards. Most groups came up with quite a number of responses. Some outcomes were most prevalent: almost all had discipline skills and generic skills such as problem solving, teamwork, communication. Other answers were less expected. For example one group had a different approach and cited employability, intellectual resilience and cross-cultural knowledge.



You can see on the next page a summary of the learning outcomes and how many times they were cited by the groups.

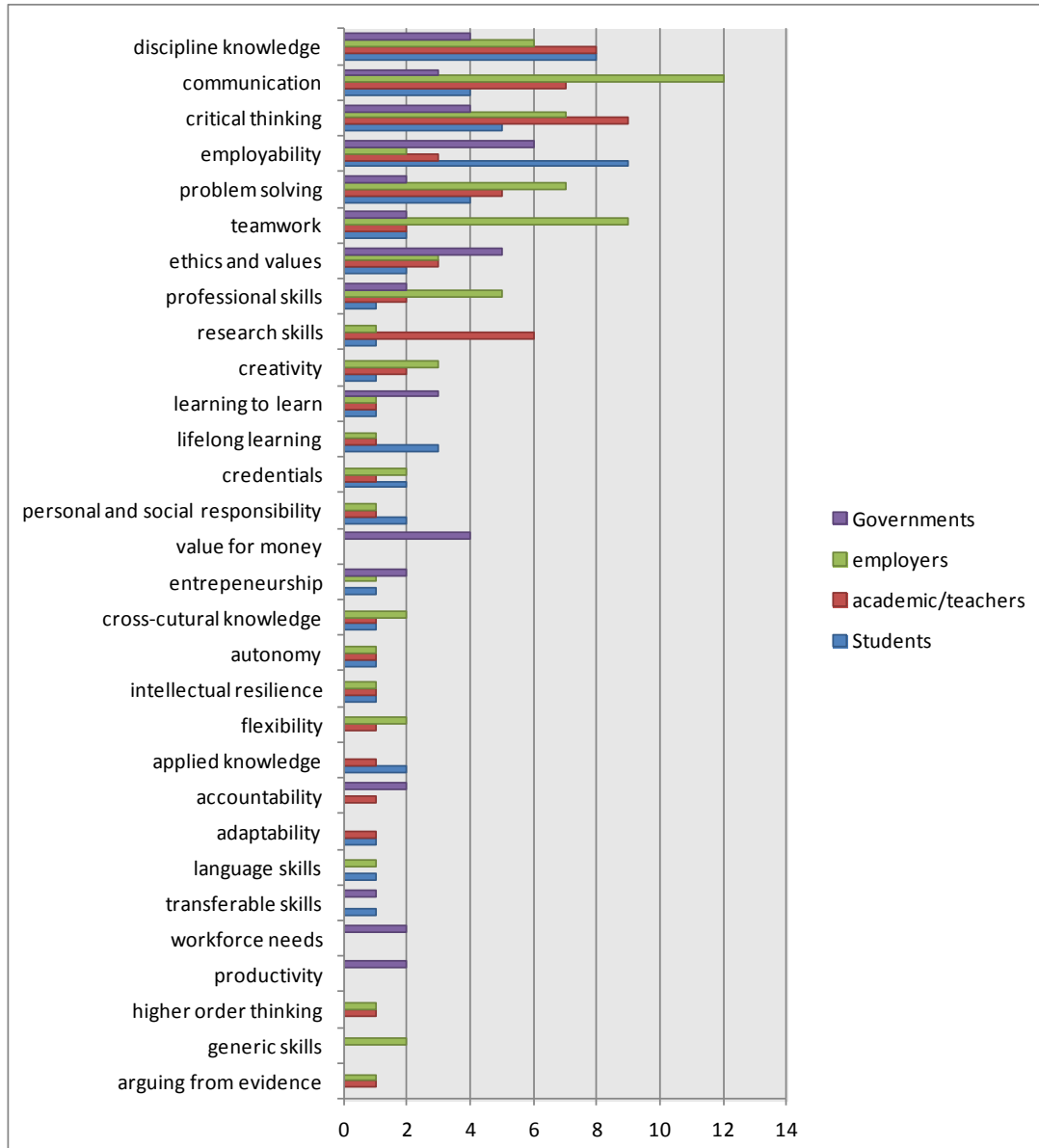
### Exercise 2a - Types of learning outcomes





## Exercise 2b: the learning outcomes by group of stakeholders

## Exercise 2b



Next we asked the groups which learning outcomes would be important depending on which group of stakeholders one belonged to (governments, students, employers and faculty/academics). The figure above lists the learning outcomes which were mentioned as important for more than one group of stakeholders or were cited more than once.

Some groups had the same list for all categories, which goes to show that dialogue is important to see if there is actually a difference. At least one group did not have anything for students. This is important to note.

The following learning outcomes were also mentioned once:

For **governments**: students learn what they are supposed to learn in the programmes, the right skills for the economy (e.g. creativity, teamwork), to get re-elected.

For **employers**: professionalism, quantitative reasoning, region/country specific, self-management, technical and soft skills, work readiness.

For **academic/teachers**: ability to ask questions, assurance and autonomy and academic freedom, cognitive skills, institutional prestige, outcomes for groups, passion and dedication to chosen field, quality feedback, resourcefulness, well-rounded graduates.

For **students**: analytical skills, competitiveness in international labour markets, deep learning rather than surface learning, global awareness, personal development, proficiency, real world application, self-realisation, social dimension, think as an individual, think of the programmes, think strategically, wide variety of experiences, work relevance of programmes, Better understanding of themselves and the world, IT skills.

### ***Exercise 3: the different measures of learning outcomes***

There are many different possible ways of assessing learning outcomes in higher education. These include approaches such as:

- Completion rates
- Student surveys
- National qualifications
- Quality assurance
- Graduate employment outcomes
- ..... and others

Each of these has a different purpose and assesses a different dimension of higher education learning outcomes. The groups were asked to choose up to three measures for assessing learning outcomes and for each measure to express how institutions could use this measure to improve learning outcomes (its strengths and drawbacks).

Some of the participants didn't like this exercise because some of the proposals were not considered true measures of learning outcomes. However if employability is considered a learning outcome then it follows that graduate employment is a measure. Further thought has to be given to the different measures and their place. The detailed feedback from the groups is included below. Again for readability we have included them under general categories.

Surveys			
Type of measure	How institutions can use it to improve LOs	Benefits	Drawbacks
<b>Longitudinal graduate survey</b>	A regular monitoring of how our graduates fare over the long term	Able to glean insights into the long-term benefits to students from their learning in school	Findings, by virtue of them being longitudinal may not be applicable to current batches of students.
<b>Student surveys</b>	Direct feedback on curriculum, staff, etc.	Customer satisfaction	More about satisfaction, often low correlation with quality
<b>Surveys/Indirect assessments</b>	Tackle weaknesses identified by students to adjust learning outcomes and change the way we teach. Create benchmark across disciplines/institutions.	Easy and cost-effective. It is consumer-oriented, student responses.	It is a proxy, has validity problem. It is only a partial piece of evidence. It is subjective. It depends if it is a satisfaction survey or an engagement surveys. Students may not respond if not engaged.
<b>Global competency: Alumni survey, Student/Faculty mobility rates and Cross-border partnerships</b>	Influence curriculum. Influence recruitment and recruitment strategy to improve attractiveness	Knowledge transfer. Improved visibility. Globalised study options.	Hard to measure. Brain drainage.
<b>Tracking the graduate system</b>	Exit survey / employer survey / alumni survey. (asking stakeholders to evaluate graduates). Results of survey need to be fed back to curriculum and academic programmes to further cater needs of the society and labour market	Relevant information to labour market, can be sensitive to changes (to improve relevant learning outcomes).	Quality of information depends on quality of questions. Gap between survey time and current needs. Could be one-sided (labour market) perceptions.

			Need to be comprehensive with multi-layered forms of information.
<b>Student survey</b>	Using results to discuss problems with faculty	Focus on learning experience. Helps institutions focus on process.	Low response rate. Not a measure of outcomes.
<b>(National) student survey</b>	The result can be used by students and student anticipates on it when they choose institutions programmes	It's a very strong instrument to get students' perspective	It is student perception therefore it is biased.

<b>Labour market outcomes</b>			
<b>Type of measure</b>	How institutions can use it to improve LOs	Benefits	Drawbacks
<b>Labour market outcomes</b>	Signal an important labour market outcome. Match with labour market needs	Signal to revive curriculum	Dependent on economic conditions. Not stable, not highly related to HEI. No carbon for entry selection. Reputation issue.
<b>Employers feedback</b>	Evaluative feedback on qualities of our graduates	Hear first-hand from employers where our graduates stand in terms of their abilities	Employers tend to be short-sighted in their assessment. They may focus more on what is important currently than what is important in the future.
<b>Employment rates / Employer satisfaction rates</b>	Aligning curriculum with job market	Increases programme relevance	Uni-dimensional education (too labour-market oriented).

<b>Graduate employment rates</b>	Give attention to areas of deficiency. Address teaching approaches/methodology. Open communication with industry.	Alignment of education and workforce needs	Making institutions beholden to industry. Risk educating students to train them.
<b>Increasing employment opportunities</b>	To become more effective and relevant	Practical measure	Diversity of local labour markets
<b>Employment / acceptance rates (professional learning outcomes)</b>	Improve the students' assessment. Improve academic offer (teaching / learning)	Close relation with labour market. Transparency. Measurable. Evaluate "over education" and horizontal mobility in labour market.	Influenced by personal/family network. Unemployment rate. Regional development.

<b>Student testing</b>			
<b>Type of measure</b>	<b>How institutions can use it to improve LOs</b>	<b>Benefits</b>	<b>Drawbacks</b>
<b>In-course tests</b>	By picking up areas in which students are weak and putting more focus on these to increase the standard.	Feedback to bring improvement	Some students and teachers may not take the opportunity to improve learning.
<b>In-course test</b>	It gives immediate direct feedback to academics and students and helps when needed to make changes in the study process.	Directly measures the performance of students, is evidence based.	Amount of time needed. Ensuring validity and reliability.
<b>in-course test</b>	Verifies cognitive learning. Could provide feedback to teachers/test writers	Easy to administer / assess. Effective scoring rubrics.	Directing learning to specific skills and aptitudes. Good test difficult to create.
<b>In-class peer review</b>	Wide-range comparative perspective. Improving quality through implementation.	Comparative perspective	Expensive, time consuming. Need of acceptance.

<b>External marking of exam papers (MCAT - GRE – LSAT)</b>	Send results back to faculty - target areas for improvement. Analyse results related to content and pedagogy. Implement any recommended changes. Monitor results.	Focuses on student learning outcomes.	Learning is much larger than these. Degree of over emphasis on text manuals. Across-culture differences pose problems for these tests.
<b>Standardised exams (M-CAT, GMAT, GRE) for post-graduate studies and professional certificate</b>	Use international exams as benchmarking exercise (internal bench-marking)	Objective: can serve as relatively objective achievement measurement.	May not be country/culturally specific.
<b>Tests (broad-based tests)</b>	Institutions could use test for accreditation (CLA). For internal improvement. To demonstrate quality.	Disciplinary benchmarking. Direct assessment of competencies.	Standardisation (by test) limits diversity. Test may not cover depth of LO we expect.
<b>Specific tests on learning outcomes</b>	Student surveys/tests on LO skills and knowledge. Provide feedback relatively rapidly.	Rapid feedback.	
<b>Accomplishment of meaningful curricular goals as assessed by the faculty</b>	Course embedded assessments as part of a continuous improvement process.	Faculty buy-in	Self-reinforced not externally validated. No comparison. Value added needed.
<b>Exit test</b>	Helps reflection on curriculum, methods. Highlights (for students) the outcomes valued. Compare with entrance test	Easy, quick, cheap. Clear reference points. Focus students' attention.	Not valid or comparable. Ignores context. No direct connection with improvement. What does failure mean? Puts focus/responsibility on students, not teachers. Increases cheating (by both students and teachers).
<b>Internal test by external</b>	Experience is the key! Application of knowledge		

<b>reviewers. Use of real life issues / problems.</b>	must be present!		
<b>AHELO (internationally normed assessment)</b>	Can provide info at end of curricular cycle of attainment of outcomes (if they are the same or similar)	Internationally developed by experts. Encourages mobility and transferability.	Could use as ranking in a wrong way. Not fine grained enough for real usefulness at the institutional level
<b>(AHELO) Cross jurisdictional externally validated assessment</b>			
<b>Testing</b>	Using international benchmark.	Clear quantified result.	Partial perspective.
<b>Discipline-specific entrance exams and graduate admissions</b>	Check how students are progressing in discipline-knowledge.	Good reflection of discipline knowledge, problem solving and critical thinking.	Do not measure soft skills or general knowledge.

<b>Quality Assurance and accreditation</b>			
<b>Type of measure</b>	<b>How institutions can use it to improve LOs</b>	<b>Benefits</b>	<b>Drawbacks</b>
<b>International quality standards</b>	Direct assessment of staff = more possibility to increase quality.	Important part of autonomy. Make university staff directly responsible for quality	No benchmark, no portable credential
<b>External QA linked to internal QA (qualifications frameworks).</b> NB: assumption that HEIs are sufficiently autonomous.	(Develop) consistency of operation/interpretation around learning outcomes should be reflected in in-course tests and exams. Ensure feedback loop with transparency internally to academics and other influences.	Consistency across programme/unit levels. Instruments like AHELO - internationally credible and referenced.	Good standards are essential - assessment rubrics, etc. Resistance to change (institutional – key individuals). Over measurement.

<b>External QA</b>	Institutions are checked by external body.	Already in many institutions (e.g. Europe)	Institution autonomy.
<b>Accreditation programme within institutions (assessment measurement).</b> <i>NB: on discussion, debate</i>	Importance of accreditation varies in regions/countries but it is an emerging trend. Reality: in some countries accreditation is one of the bases of financial support.		Quality.
<b>Internal testing of each institution (faculty) reinforced by external examination</b>	These allow them to know what is really going on within their institutions	If the measure is only internal, you can have questions about its validity. Therefore external validation (from professional bodies) is very important. It is the beginning. If these measures are not in place, how can anything else be done/compared?	How can these tests be constructed? What is the best way to measure? Timing: when should this be done? End of course/year?
<b>Internal quality assurance</b>	Internal QA measures can provide information that institution has achieved its goals such as a certain level of learning outcomes.	Supports the development of HEIs activities. Guarantees that quality is maintained at certain level (competitive factor)	Internal QA systems can be too laborious to maintain.
<b>External assessment</b>	Inviting external experts. Send products to external evaluators.	Objective analysis. Benchmarking / best practice	Perceptions of unfairness / bias. Costly / inefficient. Summative and not formative (does not focus on process).
<b>Accreditation by professional board (professional learning outcomes)</b>	Implement professional standards.	Socially recognized.	Affected by personal/family connection. Discourage students.




<b>Benchmarking and comparison</b>			
<b>Type of measure</b>	<b>How institutions can use it to improve LOs</b>	<b>Benefits</b>	<b>Drawbacks</b>
<b>Ranking</b>	Tool for strategic decisions. Reason for arguing for more resources.	Students will be happy. Benchmarking with other institutions	Manipulation of data. Rankings are only as good as the inputs. Can distort learning outcomes (or privilege some above others).
<b>Benchmarking</b>	Improve to develop these programmes.	It is useful for HEIs to review if their educational programmes and curriculum are effective or not by using the results of learning outcomes.	The limitation is that indicators of benchmarkings are used by lots of universities or not.
<b>Benchmark against national and institutional norm testing</b>	Results indicate where deficiencies/successes are. Corrective actions can be taken.	Comparability of results, supports policy development at national level. Large pool of institutions, results type. Helps keep focus on important learning outcomes rather than exams, GPA, etc.	Cost, management. Validity issues. Not performance evaluation. Design of instruments.
<b>Qualifications framework</b>	Design and revise curriculum to match desired learning outcomes	Transparency of expectations, skills, etc. Alignment of education and workforce needs/demands/expectations.	Difficult to measure. Cross-cultural differences in expectations.
<b>Practical exercise benchmarking (professional learning outcomes)</b>	Peer learning. Spread of good practice.	Measurable. Transparency.	Imitation is not always good. Reduce innovation.

## Workshop 2: What are the key challenges in developing an international measure of learning outcomes?

This workshop was designed to explore three specific challenges facing an international assessment:


- generic or discipline skills
- multi-choice questions or constructed response tasks
- getting good student response rates

### Exercise 4 – Generic Skills, Discipline Specific Skills or a blended approach



**Setting the scene**  
Exercise 4

- A key issue for designing an international assessment is what learning outcomes to assess
- One option would be to develop one assessment that measures generic skills that every graduate should be expected to learn



- Another option would be to develop to separate assessments for each discipline

Economics	Engineering	History	Political Science	Accounting
Fine Arts	Chemistry	Psychology	Physics	Nursing
Media studies	Medicine	Architecture	Religious studies	Technology
Mathematics	Logistics	Biology	Geography	Education

- A third option is to blend the two approaches

17

We asked the groups to give us their opinion on the advantages and disadvantages of assessing either generic skills, discipline specific skills or a mix of both. All the answers are provided below for each of these, grouped under sub-headings.

**Testing Generic Skills**

The challenges and costs of test development and administration

Funding instrument development - cost effectiveness  
 A challenge to do  
 Who generates/agrees the test?  
 Building consensus around the framework's methodology, content and cultural neutrality  
 Ways of testing may be more controversial in the test.  
 Difficult to agree on items  
 Ambiguous definitions  
 Difficulty in agreeing definitions of "generic skills", eg communication skills  
 Context specific  
 Not easily measured  
 Difficult to measure  
 The most interesting generic skills are difficult to measure  
 That international project has been more presumed than real  
 More challenging to assess skills (gaining agreement on which skills, assessing practical skills with written testes)  
 Difficult to define and assess  
 Elusive to define, teach and measure  
 Some skills are too broad to be measured by specific instruments like communication skills.  
 Measuring generic skills through multiple choice only  
 There are some generic skills that are hardly measurable through stardardized (multiple choice) tests.  
 Students may not have considered generic skills  
 Student buy-in  
 Engaging / attracting students for generic assessment is more difficult

*But some groups mentioned:*

Possibility of one instrument = less expensive than disciplines  
 Easier to implement internationally (less existing material to draw)  
 It is only one exam or instrument for all disciplines (if agreement is reached)

### Applicable for different populations...

Applicability in all disciplines  
 Generalisation across languages and cultures  
 One test for whole population - universal  
 Mobility/transferability across disciplines, from HEI to job market and from education to research  
 Provides a universal learning outcomes baseline for the development of discipline specific skills.  
 Value for testing in social sciences and humanities  
 transferability between nations and disciplines  
 It applies across disciplines  
 Transversal - Potential to cut across cultural differences - have good sense of what students can do  
 Internationalisation of skills, abilities  
 transferable across disciplines  
 At best, generic skills concern all disciplines in all HEIs

### ...but with some limitations

Non reflexion of cultural / language differences reduces validity for participating countries  
 Cultural issues across linguistic diversity is not timely valid  
 This would neglect diversity of student body / neglect institutional diversity  
 It is context specific  
 Cultural differences - difficult to overcome / measure  
 Culture constraints  
 Cultural differences  
 Generic skills / general education is not important in all countries.  
 No available test to measure generic skills in every country (no universal instrument)  
 Some skills cannot be measured internationally like ethics and values

### Results: complex and limited...

Harder to interpret results and use to improve learning of individual students  
 Validity, sampling, cultural bias, distance from discipline are all problematic  
 Difficult to inform faculty on improvements, shared competencies.  
 Hinders diversity and does not represent institutional priorities. Discourages diversity  
 Reduced influence by the HEIs, difficult to measure the added value of the institution  
 what is the value-added by the institution  
 Difficult to know the value added of HE in the development of generic skills  
 What is the value-added contribution of the institution? How do you measure that?

### ... but with great potential for use

promote institutions to develop their curriculars  
 more limited number of variables for comparative purposes  
 improve institution  
 accountability  
 inform faculty and students  
 larger audience for the outcomes  
 feedback is more interesting to administrators, government and non-discipline parties  
 fits in with other international processes, eg Bologna Process: Tuning Project, Dublin descriptors.  
 Gives information to HEIs to develop programmes  
 Helps to think about what is important across disciplines (multi-disciplinary competencies)  
 Helps students think about what is important to learn  
 Helps institutions benchmark  
 helps institutions focus on what is important to society  
 helps institutions think about how to use new methods to develop competencies  
 could create a continuum from PISA, PIAAC (AHELO)  
 If we reach consensus internationally about the definition of generic skills this will be great  
 Overview of disciplines in a given institution  
 Information on how to change delivery of education  
 multiple result approach  
 Benchmarking at international level  
 mobility and recognition  
 improve conceptual definition of generic skills (transparency)  
 stimulating creation of a common regional identity  
 supports mobility of students  
 Supports international mobility in global economy

### Essential skills...

Skills that HE should teach and labour market needs  
 Many of the learning outcomes are generic skills  
 Basis for lifelong learning. It is needed to cope with rapid and continuous changes  
 Required by employers  
 Important  
 All students need generic skills  
 Describe wider benefits of higher education  
 Well applicable in changing working life  
 Employability and civic engagement important in life in general  
 It measures the most important skills (cognitive) needed to succeed in professional and personal life  
 It's good we can measure such things as critical thinking or problem solving in some way

### ...but not stand-alone

Generic skills (eg employability) are largely expressed in discipline areas (particularly technical and professional)  
 Positioning vis a vis discipline specific: 1) isolating "generic skills" per se; 2) relevance to discipline  
 Lack of professional training  
 New assessment of such skills are needed  
 Evolution of "skills" needed in society = need to "revisit"  
 Lack of a common conceptual framework for generic skills assessment.  
 Higher education is constructed around disciplines. Generic skills without subject knowledge have not much value.  
 How do you separate them from a disciplinary context?  
 Neither approach is complete without the other

**Testing discipline specific skills**

Useful in the global context...

Internationalisation of many disciplines  
If the instrument is valid and reliable enough, it can provide information to enhance student mobility for example and as a marketing instrument.  
Supports the mobility of students  
Depending on subjects, it can travel (region to region)  
Employability and mobility for labour market  
Mobility and recognition  
Attractiveness of HEIs internationally

....but the diversity of local context and disciplines may create difficulties

Cultural considerations and barriers  
Diversity of students  
Variation in education systems (2 years study  $\neq$  4 years study) cannot be assessed.  
Content has to be culturally sensitive (standardisation)  
Difficult to capture specialisation needed in different regions  
family condition (equity) and national context not reflected  
Some disciplines are entwined with national politics / history / ideology  
not all disciplines can be measured internationally (physics can but history or social sciences are more challenging)  
Hinders interdisciplinarity

### Easier and cheaper...

More existing material and experience to draw from  
 Have materials from the disciplines to inform the instruments  
 Relatively easier to assess the basic and minimum outcome  
 Easier to measure than generic skills  
 Knowledge is easier to assess than generic skills  
 Easy to identify what students should be able to do  
 Easier to measure  
 Less expensive to develop

### ....but costs add up if you want to look at more than one discipline

More instruments required for multiple disciplines, may come with significant costs  
 Costs of developing tests  
 Definitions of disciplines - cannot focus on all  
 Cost, if applied to all disciplines  
 More expensive - needs more development work in each discipline  
 Difficulty of categorising some students into particular disciplines, especially with inter-disciplinary approaches.  
 Costly

### Finding consensus will require work...

Some disciplines that may be difficult to assess. How are disciplines prioritised?  
 Consensus around framework will differ according to discipline (i.e. law vs. mathematics)  
 In many disciplines it is very difficult to reach a consensus.  
 You have to design many different tests.  
 Degree vs. discipline  
 Disciplines change  
 Difficult to define requirements and competencies necessary for students to know to be proficient in a certain field.



... but the test could be more intrinsically interesting and engaging for participants...

Will engage academics because they relate to their discipline  
Easier to reach consensus on instrument (types of appropriate questions)  
Student buy-in  
Faculty buy-in easier  
could choose very international disciplines, such as economics, engineering  
higher interest of students to participate  
Easier to engage students?  
It is applicable for particular disciplines, eg engineering  
It might promote the establishment of clusters of peer HEIs to compare between themselves  
Provides benchmarks for countries in a particular discipline  
promote curricular improvement in a particular discipline  
Only way to test discipline knowledge  
Good instrument for specific knowledge  
Increased validity; may lead to more relevant samples  
more value to institutions for understanding strengths  
Allows for measurement of specialisation  
Feedback is more interesting to students and teachers  
Inherently important to departments and faculties, particularly in some fields  
Bridge between institutions and employment  
Helps quality assurance initiatives  
Highly relevant tasks to the discipline  
Provides curriculum development  
Peers define; benchmark acknowledged by profession

...provided this does not bring on an oversimplification

Affect quality of discipline because to make assessment generalised sometimes necessitates dumping the questions to make it applicable across HE systems.  
 Could limit, reduce disciplines to a minimum common core and stifle innovation.  
 Hard to recognise inter-disciplinary aspects.  
 Could freeze curriculum innovation; "lowest common denominator" approach to outcomes.  
 Selectivity: will limit the breadth of discipline coverage due to the necessary consensus.  
 Need of regular updating (curriculum / programmes) to cope with evolution of "disciplines".  
 Doesn't give a complete picture of the institution and its efficacy overall.  
 Rigidity? Flexibility needed.  
 It's not above the context yet.  
 Other regulations might bias the results.  
 Might not represent the mission of the HEIs and/or bias the results.

### ***A blended approach***

We asked the groups for suggestions on how to go about a blended approach. The majority view from the workshops was to test generic skills **within** a discipline-specific assessment.

Groups also suggested to:

- have a generic skills standardised test with discipline specific sub-sections;
- use PIAAC as a point of departure (generic skills);
- for each outcome develop a generic part and a discipline part;
- look directly at specific generic skills such as problem solving or teamwork; or
- have three components: test discipline specific skills, generic skills (abstract) and domain skills.

A few suggestions on appropriate assessment mechanisms were: case studies, a combination of different types of testing, peer evaluation of students.

The groups also had the following comments on the blended approach:

- It is important to identify which generic skills are appropriate for which disciplines.
- Many assessments are looking at outcomes that are the beginning, the minimum accepted, the goal is to go beyond that.

- The most suitable model providing that: multiple choice test must be coherent as regards context, conceptual framework, instruments and rubric.
- A blended approach can potentially maximise the advantages of both approaches and minimise the drawbacks.

*Advantages of a blended approach*

**Better feedback**

- more informative for the student
- may be a driving force to curriculum design
- indicators widened reflection on LO also among students
- as an institution it would give you the broadest and most useful range of indicators
- more valuable data and feedback

**More useful**

- more informative for the student
- more valuable data and feedback
- as an institution it would give you the broadest and most useful range of indicators
- may be a driving force to curriculum design
- widened reflection on LO also among students
- the most relevant and flexible approach
- assessment can be tailored to labour market

**You cannot have one without the other**

- you cannot do one without the other: first assess discipline specific knowledge then generic skills
- assessing generic skills require some discipline context
- generic skills are acquired through a discipline

### Comprehensive

- give more complete picture of the abilities of the graduates
- more well rounded / able students
- naturally suited to mission of university
- comprehensiveness
- more comprehensive instrument can be made than single approaches
- problem solving based in disciplines
- recognises diversity of outcomes within an institution by discipline
- could also be the best of all worlds
- generic skills within discipline may be more valid and relevant or more appropriate for the course of study
- important in particular disciplines to have particular skills mix and this differs from discipline to discipline
- multi-purpose tool
- recognises an interdisciplinary approach - many students combine disciplines

### *Drawbacks of a blended approach*

#### Limited

- world-wide applicability
- actually narrow in scope as only selected disciplines could be tested in a given year
- the cross-linguistic, cross-cultural, cross-national underlying differences may warrant less emphasis on international project and more on national
- difficult to draw results and compare them

#### Time and resource constraints

- will still be additional to what students are doing
- expensive
- longer duration

**Complex**

- assessment tools that try to test too much have less validity
- need entry measure to control for selection bias
- unclear how to adapt /translate generic skills into discipline-specific ones (e.g. lifelong learning in engineering)
- producing such a test (integrating too many approaches)
- the most difficult to design and agree upon
- It may be difficult how to blend the two approaches and which approach should hold more weight in the instrument
- differences difficult to bridge (practical)
- to fit so many national contexts the measure may lose some value
- measuring across disciplines is difficult
- possibly time consuming to produce quality "questions"
- possibly more complex to interpret the results
- could risk having worst of both worlds
- harder in developing the tests to share between discipline, good way of testing particular skills

**Exercise 5 – CRTs vs. MCQs**

The groups then considered another key challenge – how to assess learning outcomes in ways that would be useful to higher education institutions.

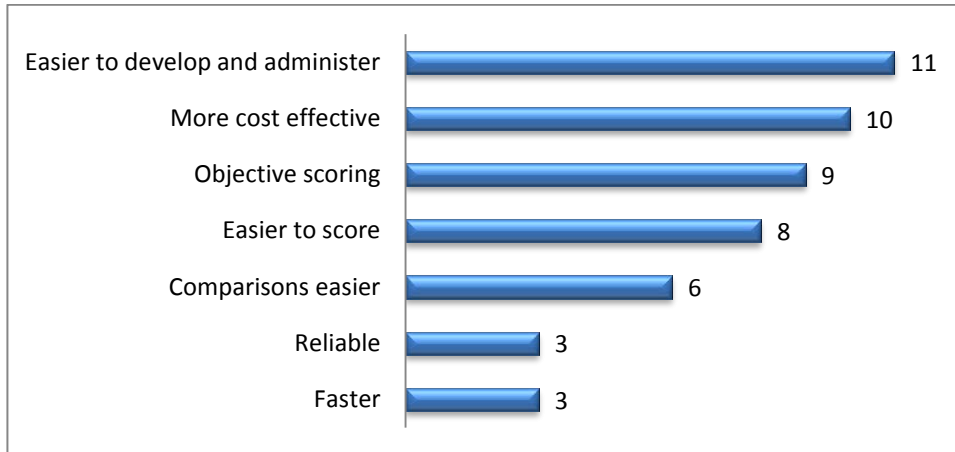
There are two broad types of questions that can be used in any sort of written test. The first approach is the multi-choice question where the student has to choose between several responses.

The second approach is a constructed response task. As the name suggests, the person being assessed needs to construct a response, usually drawing on some materials provided.

Each of these approaches has strengths and drawbacks and an international assessment could be based on one or the other or a mix of both which we asked the groups to discuss.

**Multiple Choice Questions**

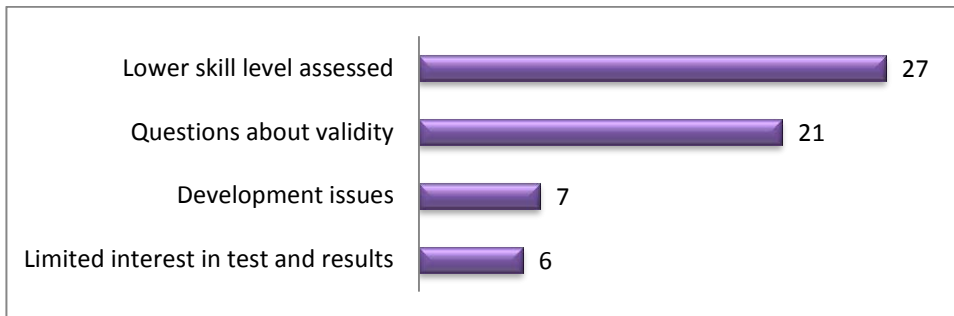
There was broad agreement on most of the strengths of MCQs:



Also mentioned once:

- Adaptable
- Applies to atomised knowledge
- Can assess higher order skills
- Comprehensive subject matter coverage
- Cross cultural validity is easier
- Easy analysis
- Easy to adapt
- E-learning is growing
- Good for discipline specific skills
- Less disciplinary bias.
- Less labour intensive
- More advanced multiple choice tests with different paths
- More tasks
- New research and technology have enabled better MC tests

On the drawbacks of Multiple Choice Questions responses from the groups still fell within a few broad categories but were much more detailed. The graph below provides a brief summary and is followed by the detailed responses from the groups.



### Lower level of skills assessed

- Cannot measure creativity in its general aspect
- Does not adequately reflect knowledge
- Respondent does not propose a solution
- Does not train student to write
- Harder to do test generic skills
- Tests recognition of answer
- Does not test application of knowledge
- Shallow perspective
- Indirect evaluation
- Unable to measure argument construction
- Simplified
- Very limiting, might not capture the whole reality
- It's hard to test the "thinking process"
- Limited opportunity to measure sophisticated skills and knowledge
- Doesn't access written skills
- Not nuanced
- Lack of depth, often surface level
- Superficial: limited in what is measured
- Can't measure some important skills (critical thinking, communication)
- Reasoning invisible
- "Teach to the test", doesn't develop skills
- Less nuance
- Narrow to curriculum
- Atomised knowledge
- Hard to capture originality of thought
- Hard to capture reasoning
- Limitations for student responses

### Questions about validity

- A limited instrument shows sometimes only luck, not knowledge
- Perhaps student familiar with patterns
- 25% probability that you are right
- Adequacy to the goal of assessment
- Reliability/validity issues
- Cultural education context does not lend itself to MCQ
- Not realistic environment
- Teach to the test
- Disliked by discipline experts
- Hard to design valid questions
- Assumes LOs can be assessed in this limited way
- Low face validity
- To set the "right" level of questions to measure one learning outcomes
- Error from guessing
- Lower (performance) predictive value
- Validity
- Less valid in most discipline at HE level
- Can learn / teach to test more, unless very large question banks
- We do not consider that MCQs can operate at the higher level in most / all subjects
- There is fear for students applying some guessing
- Multiple choice might not be the "rule" for students. This could influence student testing.

### Development issues

- Need framework
- Need years to develop good tests
- Requires sophisticated question construction
- You have to change it every time (fraud misuse)
- May work better for some disciplines than others (eg philosophy, dance, music)
- Language and translation must be clear and good
- Scoring

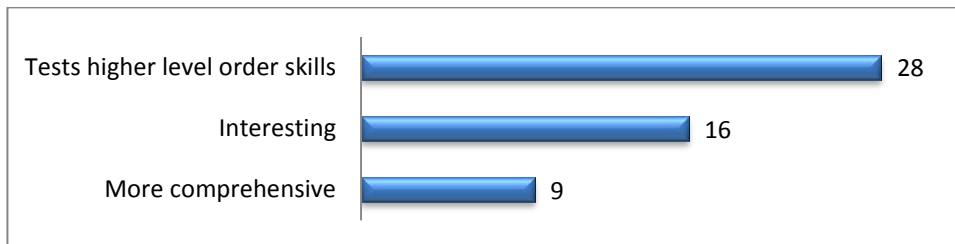
### Limited interest in test and results

- Less relevant to students, less engaging
- Anonymity
- Second best option
- Results less convincing in arguments for change
- No new ideas emerge
- Leads to unimaginative instruction



### Constructed Response Tasks

#### Advantages of CRTs



#### Tests higher level skills

- Can measure cognitive process (critical/analytical thinking)
- Student proposes solution
- Writing skills are demonstrated
- Develops creative thinking
- Taps into higher level skills (knowledge application) / generic skills
- Provides increased opportunities for students to exhibit what they have learned
- Registers mental processes
- Higher level of cognitive skills can be assessed
- More difficult "real-world" questions
- Measure writing effectiveness
- Test how to construct arguments
- How to use literature
- Tests how students think
- Better for skills and attitude learning outcomes (e.g. "think like an engineer")
- Assesses certain skills
- Allows students to apply skills and knowledge
- Depth of response and ability to analyse three outcomes at once
- Elicit deeper knowledge
- Measures critical thinking and communication better
- Provides evidence of what students know rather than what students don't know
- Possible for complex problem solving
- Integrated skills
- Captures complex / higher order thinking
- Can't learn / teach so specifically to test
- Students demonstrate their knowledge in an elaborate way
- Students integrate all their skills (writing, communication, critical thinking, decision making)
- Students are allowed to justify their answers
- Wider range of skills (creativity, etc)

### Interesting

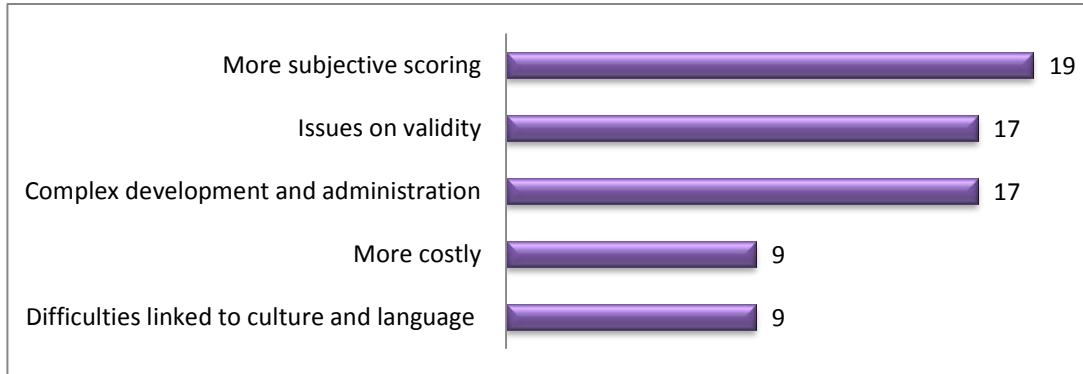
- Scoring gets teachers and faculty in important work and discussions
- More interesting for students
- Some preference among faculty
- Higher acceptance of instruments for stakeholders
- New ideas emerge
- Potentially more engaging
- Attractive for students and institutions
- More demanding (in a positive way) and challenging
- Real life decisions
- Integrates greater faculty and student engagement
- Face validity for students
- More in-depth feedback is possible
- Richness of results can lead to interesting insights
- Students like them
- Validity
- Possibility to involve students in analysing the results

### Comprehensive

- Tests discipline and generic skills
- Multidisciplinary perspectives
- Comprehensive and focused on argument construction
- More diagnostic power
- Provides opportunity for nuanced contexts
- Can measure multiple items (/skills)
- Gives more information to assess the learning
- Completeness of information
- Structured information

One group also noted that the Constructed Response Tasks may be easier to construct.

## Drawbacks of CRTs



## More subjective and difficult scoring

- Scoring: need multiple control for bias
- Complicated to train scorer and assure consistency in scoring
- Scoring is difficult
- More subjective
- Objective scoring is very difficult
- Scoring, inter-rater reliability is more difficult
- Scoring challenging
- Hard to score consequently
- Subjective
- Hand scored; subjective
- Reliability of scoring
- Expensive to mark with adequate moderation for elimination of bias
- Scoring difficulty (reliability)
- More subjective
- More judgemental in marking
- Takes longer to mark
- Much more marker training needed
- Subjectivity in scoring is very high (unless controlled by a well-deserved, featured scoring rubrics and trained scorers)
- Difficult interpretation of results (for scorers)

### Issues on validity

- Unfamiliar testing format
- Difficult to assess comparatively
- Reliability issues
- Validity
- Harder to compare
- Disciplinary bias (social science tend to higher scores)
- More measurement error
- Less reliable
- May be harder to disentangle learning outcomes
- Reliability
- Hard to differentiate
- Adequacy to the goal of assessment
- Validity
- Less reliable
- Constructed response might not be the "rule" and this would influence results
- Harder data analysis (qualitative)
- Depends on writing skills

### Complex development and administration

- At the institutional level open-ended tests are less feasible. The further from home you are the more complicated to get.
- Test design is difficult
- Time consuming
- Complicated
- More complex
- Harder to standardise
- Difficulty in standardizing assessment
- More time
- Complex to develop and to score
- Not enough data per student
- To set the "right" level of questions to measure one learning outcomes
- Time consuming
- Time for meaningful tasks may be too great for international application
- Time management (verbosity for students and scorers)
- More challenging management of assessment
- Time consuming
- Difficult/time consuming to assess

### More costly

- Higher costs
- Cost and management to develop test
- Costs and time consuming
- Labour intensive
- Costly
- Costly (resources)
- Costs more than multiple choice
- Expensive
- Could be more expensive

### Difficulties linked to culture and language

- Cultural differences
- Are the algorithms transferable across countries and languages?
- Extremely hard / impossible to remove cultural bias
- More challenging to translate
- Cultural differences
- Cultural differences can affect responses
- Very much context oriented (e.g. selection of topics)
- Translation issues
- Second language issue

### ***Exercise 6 – enhancing student response rates***

Based on the feasibility study experience, one of the biggest challenge some countries faced was motivating enough students to participate in the testing. An international assessment will only be effective if students participate. This can be a real challenge although it also depends on the country context and how it is presented within the institution.

We asked workshops participants to first individually give suggestions on how to motivate students to participate (within their national and institutional context) and then asked each group to propose three suggestions.

The individual answers are provided below. We have grouped them under general headings and noted the number of times they were cited.

36

Feedback to the students of their results  
(including discussing their mistakes with faculty)

34

Give a credit or certificate

22

Embed testing in the curriculum/existing exam

21

Clear and detailed explanation of the project and its importance  
within the global movement for evaluating learning outcomes

19

Monetary incentive (for example reduced module fees)

14

Make the test compulsory

13

Give the students their results and some benchmark

11

Communication from Universities (via posters, emails, communication  
session, ad campaign, etc.)

11

Involve students and student unions in the design of the test

10

The test should be interesting, make sense to students, clearly show how it is  
useful for their own learning.

9

Make it consequential for students (e.g. necessary for graduation or entering  
social/sport activity)

9

Gifts / vouchers / cinema tickets / other (eg priority tickets for graduation,  
parking spaces, etc.)

7

Organise the timing in a way that encourages students to participate (i.e. during  
regular class time, in a good time within academic year) - 7

5

Ally with employers in the test design or to give AHELO score employability value - 5

4

Involve faculty - 4

4

Develop an institutional "culture of assessment" - 4

- Find a way to appeal to competitive mind of students - 3
- Flexible test delivery - 2
- Included in rankings/league tables as an indicator of institutional performance - 2
- Grant to institution with highest participation rates or best results / competition between HEIs - 2
- Link learning outcomes to qualifications framework so common language which allows all stakeholders to participate (including students) - 2
- Communication structure and publication of results (links with employers) - not obligatory within the first few years of participation - 2
- Analyse the results as part of a course or project - 1
- Warning: do not make the participation compulsory (to avoid casual or insincere answers) - 1
- Incorporate in main study instrument and devise instruments in binary system - 1
- Provide some training to familiarise students with the test - 1
- Census approach - 1
- Use sample method and focus on that sample - 1
- Gaming: video game, to promote collaboration, problem solving. Video simulations as performance - 1
- Students must feel like it will make a difference - 1
- Test results relevant regarding employability (i.e. not generic skills) - 1
- Emphasize the "global" importance of this type of assessment - 1
- National mandate toward participation - 1
- HEIs in charge of administrative and organisational measures - 1
- National prizes for winners - 1

- No monetary incentives - 1
- Good feedback to institutions - 1
- Make the assessment interactive - give students the message that their input influences curriculum, testing and learning outcomes - 1
- Getting student feedback on their testing experience - 1

The group answers are provided below:

#### Embed within existing test structure or curriculum

- Generic skills is better embedded in other settings, so the assessment doesn't seem to be a "tack on"
- should be an integral part of the programme
- Building it into a regular curriculum test (but this is very hard in relation to institutional autonomy)
- Embed it into formal assessment within the curriculum - it could count towards graduation
- Embed assessment into regular assessment process (stakes involved). Graduation? (high level buy-in)
- Embed into assessment programme
- Integrate assessment into existing structures (courses, programmes, institutional assessment process)
- Integrate into pedagogy and teaching
- Integrate results into Diploma Supplement (EHEA) or Certificate of Achievement
- Integrate into curriculum - credit for taking test
- Integrate into course deliver (make it mandatory): as final year project, delivery during course
- Embedding the test in a course (obligation): ensures high rates of participation



### Provide feedback

- Giving feedback on results to students
- Provide students with their results - make it of value for them
- Making it intrinsically interesting to students (eg by giving feedback, comparing nationally and internationally)
- This assessment is part of the universities feedback system / QA system and students get feedback
- Basic motivation - it's of use to me and the institution plus some kind of feedback quickly
- Letting students know how they scored
- Feedback at the individual, department and institutional level so we can see that it is a tool for improvement
- Individual feedback to students on their test results
- To provide valuable feedback to each student on the result, how their university compares to others globally, a personalised profile/score, or certificate of participation mentioning the OECD - to enhance their CV - need to ensure employers understand this but it cannot be high risk for students
- Give individual feedback, benchmarked against peers.
- Provide results and feedbacks to students in a timely manner
- Intrinsic motivation for students - timely and useful feedback and interesting real world challenge
- Provide feedback to students on their performance
- Provide good quality feedback, local variations on usage, eg for developing statement / expression of competence across learning outcomes
- Use in context of educating students about importance of Learning Outcomes and transferable/employability skills
- Provide feedback to students: measure performance at student level to allow institutional, national, international benchmarking and to allow students to actually learn from their mistakes
- Individual feedback (with reference points)
- Enable students to get their own results and to get feedback. Don't use a sample, use all students
- Allowing comparisons between students on the same course, between different courses at different institutions, etc
- Ensure a system for feedback about the results of the test and that it will meet their expectations for positive impact
- Provide feedback to students / HEIs interesting for them

### Involvement Faculty and Students early on

- Involve faculty and students (faculty can help motivate)
- Student associations should be involved and those expectations should be used in building of the test
- Asking students what they think is relevant
- Sponsoring by professors and faculty (buy in) is very important
- Explain interest to students for improving courses: appeal to their natural curiosity; engage student bodies through campaigns / social media / lotteries
- Make it possible for students to feel ownership - engage students on research team
- Encouraging and motivating from faculty - make it a community effort
- Negotiate with student leaders to get their support for participation. Involve student organisations right from the start of the process. Try to get something out of it that benefits students if they participate, including that will improve the courses in the long run

### Communicate well

- student buy-in through PR: improve your education
- communicate value/purpose of assessment widely to create broad support for effort. Faculty who believe strongly in effort will have great influence on students.
- Awareness and communications: about AHELO: explain AHELO to students, have media campaigns, engage students clubs/unions; Recognition of high performers: reward by naming, employment awards (paid internships, etc)
- raising awareness and enthusiasm about the purposes and benefits of the test (life experience, national priority,...)
- Concentrate on motivation (aims of the test, rewards - not academic)

### Good timing and organisation

- organise the timing right (e.g. during the regular class hours)
- shouldn't be a hindrance, not conflict in any way (exam preparation, writing of thesis, etc)
- More time for the implementation phase
- Make sure testing is well organised
- Improve organisation (timing, facilities, students involvement, also in assessment, training faculty and administrators)

### Financial incentives

- International funding/grant to be given to the institution with highest participation rate or attainment
- modest compensation
- Give them a (financial) incentive
- Payment to students
- Incentives (financial - certificates)

### Awarding credit

- Allow students to participate in lieu of a course assignment or exam
- Making it part of the grading process (e.g. on DS)
- Credits (some sort of credit)
- Incentives: provide extra credits, financial, certificates, etc.
- either awarding academic credit

### Link to labour market

- Employability incentive
- Viewed as having a global importance
- Design test in a way that it is relevant to employability
- Engage students and employers in the design of the test

### Make the test interesting

- Make the exercise interesting
- students need to value the assessment
- Test must make sense: students need to see its usefulness
- Make it useful to students

**Make the testing compulsory**

- Obligatory: core part of university curricula and expectations set in study guide
- Make it consequential for students.
- Making it compulsory

**Other suggestions / comments**

- Easier in culture of assessment institution, shouldn't be a stand-alone
- Integration into local strategies of QA
- Incentives don't really work

**Workshop 3: How can we combine an assessment of learning outcomes that is useful to institutions with wider policy goals?**

This workshop was designed to identify how institutions could use the results of an international assessment to foster improvement in learning outcomes and to explore the benefits for different types of stakeholders.

***Exercise 7 - Types of data and Uses for the data***

We asked the groups to consider what types of data institutions might find relevant and useful. But since collecting data is not an end in itself we also asked them to focus on how institutions could use data to improve quality and improve learning outcomes.

There was diversity in the answers and some important new points emerged. It gives us a new set of issues and possibilities. How HEIs could use this data in improving their learning outcomes is important<sup>2</sup>.

---

<sup>2</sup> On this issue IMHE put out a guide last year on the policy levers to foster quality teaching (<http://www.oecd.org/edu/imhe/QT%20policies%20and%20practices.pdf>).

*Types of data***Data: Student performance**

- Achievement levels
- Anonymous data from the whole study
- Average and distribution of students' performance (grades)
- Average the distribution of students within the institution as a whole and in comparison to other institutions, especially similar institutions
- Data on individual student outcomes
- Data that could be used with other data to help understand patterns of student and professorate performance and the curriculum more generally
- Distribution of students in institution
- How their students compare with other students (benchmarks)
- Important data: Student performance compared to others and comparison through time.
- Individual participant data
- Individual results against benchmarks (data)
- Information on individual performance to HEI
- Internal and international comparisons of weaknesses
- Max/min analysis, percentage, median
- Mean data for performance of students
- Raw data / some analysis
- Student level data
- Student level reports
- Reliability of data and interscorer reliability
- Categorised data

#### Data: By Learning Outcome or Discipline

- Critical areas in which students exhibited low performance
- Data aggregation per outcome and per question item
- Data files (with individual student performance on each item) must be returned to institution with all data definitions
- Data indicating generic skills students have acquired from HEI and elsewhere
- Detailed information (no filters/aggregation)
- Learning outcomes sub-scores must be provided
- Raw data mapped to skills that they measure
- Results from sub-scales of questions to strong and weak areas of student performance
- Student performance in generic skills as a function of discipline/department/programme (within and between institutions, nationally, internationally)
- Sub-scores in sub-domains
- Useful to have sub-scales, e.g. to see what aspects of student performance are stronger or weaker but on a qualification or discipline basis
- What does each task measure exactly?

#### Data: At the institution level

- Information on the degree structure and teaching/learning process of other participating HEIs
- Data showing how well HEI is performing at international level
- Focus on individual satisfaction about institution
- How they compare with other similar institutions (at the sub-national, national and international level)
- Institutional data may be less threatening than local data
- Institutional level report (detailed report)
- Institutional strategic planning
- Method of comparing institutional performance
- Not student level data - the instrument is not appropriate (except to track skill performance and future employment success)
- Student support services
- Categories of data: institutional level, discipline level, national level
- Programme level reports
- Raw data on their own institution plus results/analysis of other HEI is crucial

**Data: on context**

- Background questionnaire responses
- Contextual data - student background, method of organising the studies, the type of institution
- Contextual data (parental education, socio-economic background, etc)
- Demographic and socioeconomic factors linked to performance
- Focus on Individual information about social context
- How complete can we collect a comprehensive contextual profile of students (i.e. pre-college preparation, etc)?
- Need comprehensive contextual data to make proper interpretation of data
- Performance in generic discipline learning outcomes and contextual data (could relate to staff, student ratios, etc.): capable of providing assessment over time. Multi-scale data is essential.
- Evaluation of impact of contextual factors

*Types of Analysis***Analysis: Benchmark and comparison**

- Higher level benchmark analysis
- Benchmark information
- Compare institutions by type (local categories)
- Compare results with similar departments at different institutions
- Comparison among different institutions
- Comparison by discipline and by skills
- Comparison by subject
- Comparison by subject/skill type
- Comparison with other institutions / countries
- Comparison with similar data from MOOCs
- Comparisons - a delicate issue!
- Identify link between scores and curriculum: explore curriculum differences to benchmark against similar institutions (NB: you would need results at identifiable institutional level)
- Analyse the distribution per study field
- Benchmarking won't work on most raw scores: need to take account of characteristics of populations concerned.
- Use data analysis to define and design learning outcomes which are measurable through international comparison

#### Analysis: Over time

- Compare cohorts over time
- Comparison of cohorts through time
- Time progression of institutions (performance in a, a+1, a+2)
- Year-on-year changes in performance, to check whether changes/improvements have worked
- Easier to use data to compare cohorts through time but have to know course/graduation to use effectively
- Focus on Comparison of cohorts (if the assessment is repeated)
- Item by item over time: targeted improvement of performance
- Longitudinal data
- Will need several years of data (trend data)

#### Analysis: Best practice and improvement

- Correlational analysis between high-performing institutions and pedagogical approach (for institutional analysis)
- Data analysis to identify learning activities to measure and improve students' learning outcomes
- Further analysis of "best" examples - produce case studies
- What is typical for the institutions performing well
- Analyse level of success in motivating students to perform their best

#### Analysis: Value added

- Value added by each qualification
- Value added element is important
- Value added in learning
- Value-added (or "learning gain")
- Variation of student performance (time series / longitudinal) to measure value-added
- What kind of value-added to the performance of the student (also how much value added)?
- Must figure out what is value added for one institution



### Analysis: Labour market and national policies

- Can a learning outcome include potential entrepreneurship and creativity?
- Is the programme adequately aligned with industry skills/competencies expectations at local level?
- Employability outcomes
- Data should be co-ordinated / agreed between HEIs and professional bodies
- Do correlation studies to investigate relationship between AHELO performance and work readiness (employment rate, employer satisfaction survey, alumni survey)

### Analysis: strengths and weaknesses

- Analyse results of students and faculty to identify strong/weak areas of performance
- Comparative analysis on strength/weaknesses of the programme through international comparison
- Identify strengths and weaknesses

### Other analyses

- Analyses: need to be able to construct reasons for data outcomes, in domains that can be changed within institutions, value adding to available data, cost-benefit analysis - e.g. connections between LO data and contextual over time).
- Analysis by cohort, class size
- Compare expected learning outcomes with results of assessment of acquired learning outcomes
- Different needs inside a HEI (students, faculty, administration)
- Help on how to interpret results
- Identify trends and issues
- Relationships between research performance and learning outcomes

*Data uses***Data use: Teaching and the curriculum**

- Could be used for teacher assessment
- Curriculum design and re-design
- Curriculum revision (coherence, relevance)
- Faculty development
- If you have individual student data then use for feedback
- Improve curriculum, learning experience
- In what areas is the institution deficient? What can we improve (curriculum design, re-design, programme content; pedagogical approaches)?
- Must help faculty better teaching and curriculum design
- Organise a series of workshops for faculty to discuss and review collaborative and individual teaching practices and the way forward
- Pedagogy
- Pedagogy, teaching and learning effectiveness
- Plans for correction/ improvement/ enhancement of educational process
- Reflection on teaching and assessment (feedback allows)
- Review and develop teaching and learning methods
- Review and modify curricula
- Revisit course learning outcomes in light of AHELO results
- Seek training resources so that faculty can make sense of results with sufficient nuance.
- Trigger discussion on quality teaching among faculty (content, teaching style, student expectations)
- Use data analysis to identify strengths/weaknesses in designing curriculum to improve it for better learning outcomes analysis (SWAT) pedagogical approach
- Use data analysis to identify strengths/weaknesses of pedagogy to improve teaching quality, competencies and effectiveness, to eventually improve learning outcomes
- Use in pedagogical training - enhance quality teaching
- Weaknesses of teaching/learning process improve from HEIs perform well
- Questions related to curriculum design, pedagogy and competencies acquired and relevance and reliability

**Data use: Best practice and improvement**

- Analysis of the institutions which obtained the best performance (in terms of organisation, teaching practices)
- Could validate what schools already assess
- Design of institutional policy
- Engage with the administration about how the university is structured
- Enhance the pool of better practices
- Evaluate impact of improvement, already implemented
- Gather all stakeholders to evaluate areas for improvement (student, faculty and employers)
- Genuine quality improvement
- Longitudinal analysis to see the impact on changes made with pedagogical systems
- Reflect on what possible changes could be made
- Strengths and weaknesses in performance: start to a discussion and progress inside the institution
- Target students needing additional help
- Target talented students
- Use data to highlight points of strengths and weaknesses in the educational environment
- Utilise institutional researchers and faculty planning teams to make links with quality assurance groups
- Look for explanations
- Need a collegial process to consider alongside other data to contribute to strategies for improvement
- Needs to be enclosed as part of the HEI's QA procedures

**Data use: Benchmark and comparison**

- Benchmark themselves against other faculties/disciplines
- Benchmarking at national level, institutional level, selected institutions
- Benchmarking in relation to objective standards
- Benchmarking in relation to peer HEIs at national and international levels
- Discuss the results compared with others (self analysis)
- Discuss the results with all actors
- Raises attention to comparisons across borders that have more meaning to prospective students and employers
- Encourage institutional and system-wide reflection to compare with peers (how do you find the right peers?)


### Other data uses

- Could be used for internal distribution of resources
- Information to facilitate student choice
- Interpretation of the data to overcome resistance to change
- Support for international mobility
- Assure alignment between education and employer expectations
- Use AHELO correlation studies (work-readiness, etc.) to inform career guidance

A couple groups also expressed doubts as to whether an international assessment would actually be useful to institutions and one group also pointed out that national priorities should be taken into consideration, first and foremost.

### Exercise 8

For the last of the workshop exercises we asked participants to “role play” and to answer a set of questions from different perspectives<sup>3</sup>.

		Group response card exercise 8
You are:	Your task is to tell your group:	Key points
A student	How you and other students would benefit if your institution takes part in an international assessment of learning outcomes	
Academic dean/head of department	How you would persuade your colleagues to participate in an international assessment	
Responsible for international affairs for an institution	How would an international assessment help you to do your job	
A policy official in a higher education ministry	How you would describe the benefits of an international assessment tool for institutions to your minister	
An employer	Why you would encourage institutions to participate in an international assessment	
A faculty member	How you would use an international assessment of your students to improve your students' learning outcomes	
A university president/rector/vice-chancellor	How you would persuade your faculty to take part in an international assessment of learning outcomes	
A higher education researcher	What research questions would you want an international assessment to help you address	

<sup>3</sup> One group also suggested that QA agencies / internal QA units in institutions would have been a good perspective to include in this exercise.

## Student

### How I and other students would benefit if my institution took part in an international assessment of learning outcomes:

I'd benefit if my institution paid attention to these issues  
 My institution will improve if compared to others  
 To have feedback about what I learn  
 Hold professors accountable. We need better quality education, more value on education  
 Might promote career  
 Reputation - good or improving  
 Improvement of quality  
 We want an international competitive education  
 Pressure towards the university leaders to improve results  
 Enhances and makes international mobility more transparent  
 You want to know how your institution/faculty is doing internationally  
 More feedback  
 Benchmarks with other students/institutions  
 Might motivate students to do well  
 Learn from other institutions and understand value where we stand in the world  
 Disagree: western capitalistic attempt to suppress diversity // Agree: want qualification to be recognised overseas  
 Use this assessment to mobilise myself so that I can expand my opportunity to get a job anywhere I want  
 Reassurance of quality and potential for institutions to make improvements  
 Encourage internationalization across a broader front  
 Possible interaction with other students  
 Constructive competition between faculties  
 Could be quality label to assist employability (if performance is good)  
 Get paid (incentives)  
 Increase institutional reputation and degree value  
 Help students find out if their education gives what it promises, before they start  
 Compare my results with other students.  
 Benchmarking of institutions  
 I want to know how competitive I am on a global scale  
 Life and work skills; discipline skills; leadership, languages  
 Understanding individual competitiveness and knowing what the global market values  
 Help prove/validate value of my degree  
 Show worth of my degree in a global context  
 Might help bring improvements for students coming later

**Academic dean /  
head of department**

**How I would persuade my colleagues to participate in an international assessment:**

Challenge: we think we are good then let's prove it  
 Focus on you professors: international opportunities, reward system: research support, ICT, etc.  
 Curriculum improvement improves teaching.  
 Marketing/fundraising (not in Germany)  
 Responsiveness to students  
 It will help us improve our academic performance and when getting accredited we will be in a better position  
 Share information on a need of international assessments (mobility issues as well)  
 Participation gives you a chance to benchmark and improve your programme  
 If teaching is important research could demonstrate good teaching  
 Important to prepare global graduates: need to benchmark globally  
 Important for QA and marketing  
 Recognition of qualifications  
 I would present this project to my colleagues as a tool for international visibility to promote joint research, as part of an internationalisation strategy  
 Engagement in design: emphasize improvement  
 Benchmarking, comparative  
 Engage in student motivation  
 What's in it for institutions?  
 Weak results can be argument for resources?  
 Stick/carrot  
 Make quality more visible  
 Share information about student performance  
 Ranking department / HEIs  
 Comparability and benchmarking  
 Feedback on department outcomes?  
 M+E  
 We believe we are good - so this can be demonstrated  
 Help demonstrate to institution that investment is worthwhile

**Responsible for  
international affairs  
for HEI**

**How an international assessment would help me do my job:**

Develop exchange programmes: identify sympathetic programmes  
Improve diversity of student body, especially generic skills  
Organise system of faculty and staff exchange  
Would help me choose partner institutions, higher rank or similar  
Puts my institution in international context. Helps me understand other institutions, where to form alliances.  
Use it as a marketing tool  
Make student choices transparent, gives a common language  
Use to recruit international students  
Convince partners of quality  
Connection to mandate to globalise  
Practice what we preach  
Good for recruitment  
Employability abroad  
Use assessment to pick out the best ones  
Communication with other international HEIs  
Help situate my institution among global world  
Benchmarking purposes, institutional positioning in a globalisation context (funding from international sources)  
New indicator/ tool focused on education (not just research)  
Shows that we are internationally active  
Have international students and staff  
Could help improve any weaknesses

Policy official in a higher  
education ministry

**How I would describe the benefits of an international assessment tool for institutions to my Minister:**

Good for attracting international students and enhancing competitiveness  
 This will help us identify the strength and weaknesses based on evidence and direct the reform of our educational strategies and policies.  
 Accountability of HE. We need tools for that.  
 It gives information on how faculty is doing and how the national "system" is doing on an international scale.  
 Potential tool for distribution of funds, allocation of positions, etc.  
 Use for ranking institutions / countries internationally  
 Becomes QA for some countries that don't have a system now.  
 We can discover what is working well and not working and how we can improve.  
 Instrument for institutions to benchmark themselves and also useful to know how national performance measures up.  
 Need to have publicly available data for public institutions  
 To clarify international status (strengths/weaknesses) to decide the directions  
 To set more funding for education  
 Measurable benefits for the investment  
 Only one measure... but you'll be able to engage with students that education is valued by employers  
 Risk if outcomes is not good  
 Test identifies areas for improvement. Can focus our strategies.  
 The test provides with employment/job-related data  
 Connection to international policy making, debate  
 Define institutional mandate  
 International benchmarks may improve quality  
 Don't have to develop national tests  
 International reputation  
 International co-operation  
 Benchmarking not against other comparable institutions  
 Coherence with national strategy  
 Benefits of benchmarking for policy making  
 International co-ordination  
 The results of the assessment will help improve graduate performance. Thus the HEIs will be able to deliver globally competitive graduates who can contribute to raise the national economy.  
 Local performance in global context (benchmark, comparison)  
 Widen job market for graduates  
 Justify investment in HE  
 We believe we are good - so this can be demonstrated  
 We have to participate in international work if we want an international reputation.  
 But if we participate we have to take it seriously  
 Ask first feedback from HEIs on the results of assessment and then report to Minister



## Employer

### Why I would encourage institutions to participate in an international assessment:

Increased global competition means we need to develop talent that is job ready  
 We need to bridge the gap between our worlds. We need workers with skills  
 It depends on the type of employer. SME less interested while global/international employer is more interested  
 Need to have employees to manage world-wide industries. Need to have technical or general skills.  
 Information to employers of students' learning gains / outcomes  
 Brings assurance that graduates have the requisite skills and abilities  
 Need to know we are giving students the right kind of qualifications. Companies want value for money.  
 We need to understand the quality / capacity of the graduate of the institutions so use the data to identify the competency (benchmark for local institutions against other countries)  
 Companies doing international trade need to know the competencies of graduates for employment. Evidence is critical. Looking for evidence of international level of performance. Want to see performance of students and institutions.  
 Responding to global talent needs  
 Development training  
 Partnership with local institutions  
 Clarify local labour supply  
 If of a multinational company I'd want this (maybe not in a small country). Or perhaps part of the assessment  
 To encourage institutions to develop graduates I would like to hire - a minimum standard  
 Develop adequate skills for international competition  
 Transparency of qualifications  
 Economy is controlling the world  
 HEIs need to be benchmarked and to fulfill the requirements of the labour market  
 Need qualified graduates  
 How particular institutions are preparing graduates for the 21st century workplace in selected fields of knowledge  
 Shows level of preparation for labour market  
 Graduates compete in international labour market  
 Get global "good practice"

## Faculty member

### **How I would use an international assessment of my students to improve their learning outcomes:**

Review degree, plan curriculum design.  
Review content and pedagogy and exams  
Develops exams that focus on outcomes of the course: use assessment tool samples to develop assessment  
To know of the level of performance of my students, discuss with colleagues, review teaching styles  
I would discuss this with colleagues and find a common strategy then identify lesson learnt on what works.  
What?! Can we avoid it?  
Look at weaknesses and how to improve performance  
To reflect on my teaching to improve / for better student learning  
Motivate students. Unique opportunity of unmarked test that will demonstrate how good we are - you will get outcomes with no disadvantage.  
Once we have the outcome we will be able to improve the course.  
Stick/carrot  
Individual approach: discuss with faculty leaders  
Identify gap and success ingredients to work for improving course, teaching and learning  
Strength / weaknesses analysis (individual students, as class) and engage in discussing areas of improvements  
Engage HE researchers to help better understand how to improve teaching methods  
Develop understanding of what it can and can't tell me  
Try to identify which part of curriculum or process needs to change (or faculty)

## University president, rector or vice chancellor

### How I would persuade faculty to take part in an international assessment of learning outcomes:

Focus on young faculty  
 Establish assessment as part of institutional community  
 Hold workshops - facilitate usage  
 Competitions within faculty  
 Must promote our international reputation. Prove that we are good and if we are not good fix it.  
 We need to know how we are performing compared to others  
 To see our institution position (benchmarking)  
 "Our university has decided to take part in this. You had your saying then, now you do not"  
 Use the power of a rector's authority  
 Quality evaluation of successful activities  
 Sharing  
 Similar to dean. Could be used for resource allocations, rewards, recognitions  
 How have students in my institution performed against those in others?  
 Learn where university stands and make improvement. Use data to ask government for more funding  
 Plan to offer faculty members motivation / incentives to enhance teaching, use data for monetary incentives, to write better student learning outcomes plans.  
 QA regime for institution  
 Spur international exchange of knowledge and faculty  
 Increase graduate studies by international co-operation  
 Maintaining competitiveness at local/international levels begins with knowing how/where we stand from other HEIs  
 High level of performance is final result of faculty  
 Visibility/international positioning (competitiveness)  
 Quality of teaching  
 Understand relevance of teaching with labour market needs  
 Link to mission - and put research first  
 To have global reputation we must participate in international work and research  
 To learn about and implement needed improvements  
 Help to get investment in the university

## Higher education researcher

### What research questions I would want an international assessment to help me address:

Performance of system compared with other countries, such as PISA, PIAAC is an assessment of our possibilities.

Investigate social dimension of HE outcomes. Link to accountability agenda.

How we compare with research conducted and produced in other universities

Learning gain - where does it come from?

How results relate to inputs e.g. pedagogy, faculty preparation, learning environment, resources.

What dimension or aspects may be stronger or weaker to know where to place resources or attention

Difference if students get formative feedback vs. summative assessment results

Similar outcomes for different types of students

Focus on larger variables of student performance, rather less trivial. What accounts for spread around the average rather than the average itself

Performance related to student, institution and particularly system characteristics

Come up with research question

Come up with methodology to effectively use the data

For paper writing for career development

Another point or validity anchor added value to the field

How the HE in my country stands compared to other countries (resources, cost-effectiveness, education effectiveness) to deliver the same learning outcomes.

How to initialise research (find relevant literature, etc.)

Technical report skills

Analysis of data

How to improve effective teaching?

Depends on what you want it to do - e.g. league tables or feedback to students or enhancement of learning opportunities.

## 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](http://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 third volume of the feasibility study report presents further insights on the Value-Added Measurement and the proceedings of the Conference which concluded the feasibility study.

It follows a first volume on design and implementation which was published in December 2012 and a second volume on data analysis and national experiences published in March 2013.

#### Contents

Chapter 10 – Report from the Expert Group on Value-Added Measurement  
Chapter 11 – Conference proceedings

More information on [www.oecd.org/edu/ahelo](http://www.oecd.org/edu/ahelo)  
Contact us: [ahelo@oecd.org](mailto:ahelo@oecd.org)