Introduction

On February 2 and 3, 2017, CAE, in conjunction with the OECD, hosted a conference at Homerton College, Cambridge University in Cambridge, UK. The purpose of the meeting was to discuss the use of CLA+ for making evidence-based policy decisions pertaining to use-inspired research in higher education and in the assessment of higher education learning outcomes. In addition to presentations by the Council for Aid to Education (CAE) and the Organisation for Economic Co-operation and Development (OECD), several colleagues who have collaborated with CAE shared their experiences with using CLA+ and other instruments measuring higher education students’ learning outcomes.

Session 1: Opening Remarks

What’s the Purpose of Assessing Higher Education Learning Outcomes?

Dirk Van Damme, OECD

Assessing student learning outcomes at the tertiary level is necessary and essential. However, the OECD Education Policy Committee has decided not to move forward with a full-scale Assessment of Higher Education Learning Outcomes (AHELO) main study at this time. Nevertheless, the OECD supports CLA+ International as an alternative method to measure generic student learning outcomes at the tertiary level and will help to facilitate and further conversations of interested parties. The CLA+ International initiative is not intended to act as the Programme for International Student Assessment (PISA) or to replace AHELO for tertiary education. Although the debate over measuring students' knowledge and skills in the disciplines versus generic skills persists, the Memorandum of Understanding between the OECD and CAE indicates OECD's support of countries pursuing generic skills assessment through the implementation of CLA+.
WHAT’S THE PURPOSE OF ASSESSING HIGHER EDUCATION’S LEARNING OUTCOMES?

Dirk Van Damme

OECD/EDU/IMEP
WHAT IS IT?
Objective: test whether it is possible to

- measure what undergraduate students know and can do upon graduation …
- … across diverse countries, languages, cultures and types of institutions

For

- Generic skills
- Economics
- Engineering

Assessing Higher Education Learning Outcomes (AHELO) Feasibility Study (2008-12)
A range of geographic, linguistic and cultural backgrounds involved

Generic Skills
- Colombia
- Egypt
- Finland
- Korea
- Kuwait
- Mexico
- Norway
- Slovak Republic
- United States (CT, MO, PA)

Engineering
- Abu Dhabi
- Australia
- Canada (Ontario)
- Colombia
- Egypt
- Japan
- Mexico
- Russian Federation
- Slovak Republic

Economics
- Belgium (Fl.)
- Egypt
- Italy
- Mexico
- Netherlands
- Russian Federation
- Slovak Republic

Observers
- Bahrein
- Brazil
- Saudi Arabia
- Singapore
Quick facts on AHELO Feasibility Study

- Data collection from February to June 2012
- 17 countries involved in 25 strand replications
- Data collected from over
  - 23,000 students
  - 4,900 faculties
  - 270 institution coordinators
- 1,000 test sessions and 20,000 computers involved
- Scoring completed in June 2012
- Analysis of results and findings
- Final reporting 2012-13
The main **conclusions** of the meeting were:

– The revised proposal should frame the assessment of higher education learning outcomes around **transversal skills**, and build any disciplinary assessments and contexts around those.

– Individual student results and feedback will be critical to institutional and student engagement and, therefore, the assessment should **extend to all eligible students** of participating entities, rather than a sample only.

– To report results from the assessment both in **absolute terms** (bottom-line results) as well as **relative** to the socio-economic and institutional context of individuals and universities (analytic value added).
The main **conclusions** of the meeting were:

- To devote greater attention to the development of intermediate outputs (frameworks, instruments and methodologies), not only to ensure that participating entities obtain short-term value for their investments, but also to facilitate the important dialogue with the academic community and other stakeholders.

- The AHELO Main Study should be governed by national or subnational government authorities, but the governing body should be complemented with a strong Academic Advisory Group from among participating institutions that would oversee much of the substantive development of the assessment. The governing body would be made up of those countries or subnational entities that contribute to the development and financing of the project, while recognising that not all participating entities may be able to implement the assessment from the outset.
The AHELO Scoping Paper has been revised accordingly and published as EDU/EDPC(2013)17/REV3

A new funding scheme has been developed.

Countries are invited to express their willingness to participate in Year 1 (2016) of the AHELO Main Study on the basis of this proposal and to do so before the end of May 2015.
SHOULD IT BE DONE?
Higher education is a rapidly expanding and globalising system with extremely low levels of systemic transparency.

Lack of internally generated transparency leads to various forms of external mechanisms, most importantly rankings and consumerist approaches to quality:
- Often not providing right incentives to improve teaching and learning.

Various national attempts to measure learning outcomes in higher education, but lack of international, comparative approach ≈ PISA.
Drivers

1. Globalisation of demand: what are students actually buying?
2. Outcomes and impact: what do employers and wider society get?
3. Doubts on the value of HE qualifications: selection, credentialism or learning gain?
4. Growing dissatisfaction with the way rankings are ‘ordering’ the HE system and distribute rewards and incentives
1. Globalisation

- Explosion of global demand in emerging countries
  - In what exactly are the emerging middle classes investing enormous resources?

- Global imbalance between demand and supply of supposed academic excellence
  - Imbalance which cannot be met by international mobility (4.5m now, expected to increase to 7.2m in 2020) and e-learning/MOOCs
Global expansion & redistribution of qualifications
Global distribution of tertiary educated 25-34 y-olds in 2013 and 2030

2013:
- China, 17%
- United States, 14%
- India, 14%
- Russian Federation, 10%
- Indonesia, 4%
- Japan, 6%
- Brazil, 4%
- Korea, 4%
- Germany, 2%
- France, 2%
- Turkey, 2%
- Canada, 2%
- Poland, 2%
- Saudi Arabia, 1%
- Australia, 1%
- Italy, 1%
- Other, 6%

137 million 25-34 year-olds with tertiary education

2030:
- China, 27%
- United States, 8%
- India, 23%
- Indonesia, 5%
- Japan, 3%
- Russia Federation, 4%
- Brazil, 5%
- Korea, 2%
- Mexico, 2%
- Germany, 2%
- France, 1%
- United Kingdom, 2%
- Turkey, 2%
- Canada, 1%
- Canada, 1%
- United Kingdom, 2%
- Saudi Arabia, 3%
- Australia, 3%
- Other, 6%

300 million 25-34 year-olds with tertiary education
Global distribution of academic graduates and academic excellence

Share in academic graduates 2010

- United States, 13.7%
- China, 17.8%
- Russian Federation, 10.9%
- Japan, 6.9%
- India, 11.4%
- Other, 11.7%
- Turkey, 1.7%
- Brazil, 3.0%
- Canada, 2.1%
- Spain, 2.2%
- Indonesia, 4.3%
- United Kingdom, 2.9%
- Mexico, 3.0%
- Korea, 3.9%

Share in academic excellence THEWUR 2012

- United States, 43.2%
- United Kingdom, 13.8%
- Other, 8.4%
- Hong Kong, 2.0%
- Korea, 2.2%
- Sweden, 2.6%
- Japan, 2.5%
- France, 3.0%
- Switzerland, 3.5%
- Australia, 4.3%
- Canada, 4.3%
- Germany, 4.3%
- Netherlands, 6.0%
2. Persistent high impact of HE on economic and social outcomes

• Economic outcomes
  – Employment
  – Earnings

• Social outcomes
  – Social capital, interpersonal trust
  – Health
  – Political participation
  – Volunteering
  – Anti-social behaviour
Employment benefits very significant
Employment rates among 25-64 year-olds, by educational attainment (2012)
Percentage effect on earnings of...

...Secondary Education

...Tertiary Education
Social outcomes – interpersonal trust
Proportion of adults reporting that they trust others, by educational attainment (2012)
Social outcomes – trust in political institutions

Individuals with higher level of education more likely to believe they have a say in government

- Below upper secondary education
- Upper secondary or post-secondary non-tertiary education
- Tertiary Education
3. Knowledge & skills value of higher education qualifications

- What do we know about the value of higher education’s qualification in terms of knowledge or skills?
Formal education vs skills

OECD average

Quantiles

Years of Education

Proficiency
Formal education vs skills
Numeracy scores of tertiary educated adults of 25-34y old
Numeracy equivalent of tertiary qualifications
Proportion of 25-64 year-olds scoring at PIAAC numeracy level 4 and 5, by educational attainment of the population (2012)
Literacy equivalent of tertiary qualifications

Japanese high school graduates have literacy skills comparable to those of Italian tertiary graduates.
4. Structure and hierarchy in HE and the associated incentives system

• Using the THEWUR database on the Top 200 universities, how dynamic is the HE system?
  – Teaching – research – citations

• Current input and reputation measures do not provide the right incentives for improving or innovating teaching & learning
Institutional moves in the THEWUR

Absolute moves

Teaching
Poly. (Teaching)

Research
Poly. (Research)

Citations
Poly. (Citations)

International
Where are the top20 performers?

- Teaching
- Research
- Citations
- International
- Relative share of top performing uni (any scale) per country (right axis)
Conclusion: what are the value-propositions?

- Is it possible?
  - To improve our understanding of what students actually ‘learn’ in higher education
  - To exchange reputations with empirically grounded observations of quality of teaching & learning
  - To gradually transform the field on which credentials are traded into a more level playing field
  - To provide better information to students and employers about the quality of teaching & learning experiences
  - To develop feedback loops to improve teaching and learning
  - To reward and incentivise institutions that significantly improve their teaching & learning environments
  - To re-confirm the value of teaching as part of the university’s mission next to research
CAN IT BE DONE?
Disciplines or generic skills

• Strong preference among interested countries for generic skills
  – Generic academic skills are at the core of the mission of universities
  – Specialised, disciplinary knowledge and skills is more different
  – Overlap with established systems or practices on disciplinary level
  – Strong preference among employers as well
Various scales

- How to respect and value institutional heterogeneity and diversification?
- Developing various scales?
  - Research skills
  - Complex reasoning
  - Creative thinking
  - Problem-solving
  - Social and emotional skills
  - Communication skills
  - Employability and entrepreneurship skills
  - Etc…
Who benefits: student, institution, country?

- Orientation towards assessment with clear benefits for students
- Countries want a census-type of assessment
- Necessity to provide something of value to universities
  - Balancing accountability and improvement functions
Absolute or value-added

• Real value-added assessment considered to be the best option, but at the same time unrealistic

• Two approaches
  – Absolute scores at individual, institutional and country level
  – Analytical value-added scores based on regressions to background variables at institutional and country level
Thank you!

dirk.vandamme@oecd.org
www.oecd.org/edu/ceri
twitter @VanDammeEDU
Initiative Conference Opening Remarks
Roger Benjamin, CAE

The focus of this conference is on use-inspired research in tertiary education. A number of presentations will illustrate the practical utility of using CLA+ assessment results to shed light on the test’s predictive validity for success in the work place (Zahner & James, 2016), use of CLA+ certified badges for job placement (ANVUR), strengths and weaknesses of Mexican universities (Rosas, 2013), and leveling the playing field from university to career for high-ability students from less-selective institutions.

There is growing interest in international, comparative assessment for a number of reasons. University graduates and employers no longer stay within the borders of the country in which they were born. Globalization creates the need for new international and universal metrics to reduce information asymmetries and to give graduating university students additional credentials they for differentiating their skill levels to employers who are looking for better information to identify potential job candidates. Moreover, all countries and large sub-national consortia of universities can benefit significantly from enlarging the sample of students and institutions they use for comparisons of how well their institutions are doing with similarly situated institutions internationally. Furthermore, placing an international focus on tertiary learning outcomes will produce greater interest and support of policymakers everywhere.

In addition to the CLA+ International initiative, three other initiatives demonstrate growing interest in educational assessment. The first is Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe (CALOHEE)—which is funded by the European Union—and is focused on developing assessments for five academic disciplines. The assessments are to be developed by faculty with no involvement from measurement scientists. This initiative does not intend to engage in inter-university comparisons. Second, Performance Assessment Learning (PAL), an international center for development of new types of performance assessment, has been launched in Berlin by Dr. Olga Troitschanskaia and her colleagues at KoKoHs. The center intends to bring researchers from government agencies, universities, and testing organizations together to work collaboratively. Finally, the Measurement of College Learning project, led by Dr. Richard Arum, is focused on the development of assessments for six core academic disciplines. The faculty in these disciplines is to take the lead in developing the conceptual frameworks and essential competencies with the participation of measurement scientists who will be tasked to ensure reliable and valid assessments are developed.

The Case for Development of an International Assessment of Generic Skills Such as CLA+

First, the case for generic skills to be the basis for an international assessment is strong. Generic skills are independent of academic disciplines. There is evidence that they can be improved through teaching, and can be measured. Unlike academic disciplines’, general transversal cognitive skills are invariant and unchanging over time. Finally, the generic skills of graduating college seniors are the joint product of all the courses they take and the experiences they have at their university. That means the institution is the appropriate unit of analysis for assessing generic skills.

If the goal is to launch an international, comparative assessment of utility to universities and ministries of education, the best course is to start with an assessment protocol that is recognized as a reliable and valid instrument in multiple national contexts. Then, transform the protocol into a truly international test while working collegially with colleagues involved in the CLA+ International initiative. Without a credible generic skills protocol to build on, the development of a sustainable, agreed-upon international-based framework for essential concepts and competencies is problematic.

The Three Principles of the Use-Inspired Assessment Approach

1. The instrument must be reliable and valid across institutions within and across national boundaries (see papers on reliability and validity at cae.org)
2. The results of the assessment must be actionable:
   • Aligned with and lead to improvements in teaching and learning
   • Provide benefits for students, institutions, national consortia, and countries
3. Governance

- Participants will own their data
- Representatives from participating countries/consortia/institutions will recommend (with the assistance of participating CAE measurement scientists) the essential concepts and competencies, test content, and new forms of tests
- Participants use the test for evidence-based decision-making

The Rationale behind the Number of Participants in the First Test Administration of CLA+ International

It is essential to include a sufficient number of participants to produce a set of findings that can be critically evaluated. If the first test administration is successful, we will have confidence to grow the number of participants significantly. We desire between three and five participants, defined as ministries of education or sub-national consortia of universities. This amount of participants will give the initiative a sufficient critical mass of participants who will provide test results that each participant can use for a variety of purposes.

The Goals of the Meeting

- Stimulation of candidate policy problems for participants to focus on
- Raise questions about the goals and strategy planned for the CLA+ International initiative
- Establish the calendar—the steps and the timing of each in the upcoming test administration
Initiative Conference Opening Remarks
Roger Benjamin  |  February 2, 2017  |  Cambridge, UK
Pasteur’s Quadrant:
Use-Inspired Research (UIR) in Tertiary Education

• Pasteur’s Quadrant is from Pasteur’s life narrative in which he pioneered use-inspired basic research.

• At historical junctures decisions created objective tools of analysis for agriculture, health and national security to be used for policy applications. All tools developed are based on value system of science.

• Increased transparency and peer review are needed for higher education. Tertiary education should be next for UIR. Our focus here is on international assessment. Cognitive science, economics, data analytics, education technology, assessment, and teaching and learning form basis for interdisciplinary focus on learning outcomes.
Why Are International Comparisons Important?

- College graduates and employers do not only stay within the country they started in---globalization creates the need for new global metrics to reduce information asymmetries.

- All countries or national consortia of private universities can benefit significantly from enlarging the sample of students and institutions they use for comparisons across countries.

- Placing an international focus on tertiary learning outcomes will produce greater interest and support of policy makers for higher education everywhere.
Growing International Assessment Initiatives Include

• CALOHEE - funded by European Union, 15 countries with 75 universities, focused on five disciplines, led by Dr. Robert Wagenaar, University of Groningen. Assessments to be developed by faculty with no current involvement of measurement scientists.

• PAL, proposed international center for development of new forms of performance assessments led by Dr. Olga Troitschanskaia and colleagues. The Center would bring researchers from universities, testing organizations, and possibly government agencies to work collaboratively.

• Measurement of College Learning led by R. Arum focuses on six core academic disciplines—faculty led with measurement scientist participation.
What is the Case for the Generic Skills Initiative?

- Generic skills are independent of academic disciplines and can be measured and improved through teaching.
- Unlike academic disciplines, generic transversal cognitive skills are invariant, unchanging over time.
- The generic skills of graduating seniors are the joint product of all the courses they take and the experiences they have at the university.
Why Start with Existing Generic Skills Protocol Such As CLA+ International?

• The best course is to start with an assessment protocol that has gained approval as a reliable and valid instrument in multiple national contexts and transform the protocol into a truly international test working collegially with colleagues involved in the Initiative.

• Without a credible existing generic skills protocol to start with, development of an sustainable, agreed upon, international-based framework for essential concepts and competencies is problematic. There is no Initiative for potential participants to join.

• The MOU references the importance of investigating new approaches to measurement and additional skills to measure.
The 3 Core Principles of the Use-Inspired Assessment Initiative for Assessment

1) The test must be reliable and valid across institutions within and across national boundaries (see papers and monographs about reliability and validity of CLA+ at cae.org).

2) The test results must be actionable:
   - Aligned with and lead to improvements in teaching and learning
   - Provide benefits for students, institutions, national consortia of institutions, and countries.
Principle 3

- Participating countries, national consortia of universities, and institutions must own their data.

- Representatives from participants at the institution, national or international level should recommend the essential concepts and competencies, test content, and new forms of tests assisted by participating measurement scientists.

- Assessments that are seen to be authentic by users and reliable by measurement scientists are then implemented.

- Participants use the test for evidence-based decision making. Ideas for test improvement is an ongoing process.
The Rationale Behind Number of Participants in this Test Administration

• Include a sufficient number of participants to produce a set of findings that can be critically evaluated.
• If the first test administration goes well, we will all have confidence to grow the number of participating countries and additional sub-national systems of universities significantly.
The Preferred Number of Participants

• We desire between 4 to 6 participants, defined as ministries of education or sub-national consortia of private universities for the initial test administration. Why?
  – This will give the Initiative a sufficient critical mass of participants to produce testing results that each participant can use and provide a large enough sample for evaluation purposes.
  – If the first test administration is successful, we will have the confidence to grow the membership in the Initiative significantly.
Goals of the Meeting

• The Agenda: stimulation of candidate policy problems to focus upon by the participants.
• Raise any questions about the goals and strategy planned for the Initiative.
• Establish the calendar, the timing, of the steps in the upcoming test administration.
Measuring Learning Outcomes at the University: The Italian Experience with TECO

Alberto Ciolfi - ANVUR

Law 76/2010, among other tasks, requires the Italian National Agency for the Evaluation of Universities and Research Institutes (ANVUR) to assess not only the processes and the inputs of the national higher education system, but also “the quality of the results and products of management, teaching and research activities... (Art. 3, c.1)”... constitute also object of evaluation the learning outcomes... (Art. 3, c. 2). Subsequently, Law 240/2010 (Art. 5, section 3) and Legislative Decree 19/2012 initiated the process leading to the integrated Self-Assessment, Periodic Assessment, and Accreditation system (AVA). Among other things, this system provides institutions and single programs with quantitative indicators relative to students' input characteristics (e.g., secondary school type, graduation examination result); students’ careers (e.g., dropout rates, time for graduation); and job opportunities for graduates. However, none of these surveys provides a direct measure of the outcome that can be used to compare institutions and/or degree programs within the same institution.

As reported in the new guidelines for the AVA process (December 2016), ANVUR will update the set of quantitative indicators according to the results of experiments conducted on the measurement of the learning outcomes (generic and/or specialist), as well as results from the survey on students' opinions. Thus, in the future, the actual results achieved by university students in terms of both specialist and generic learning outcomes must not only be compared with the expected ones, but also certified, and the relative indicators be considered for the assessment and accreditation purposes.

Starting in late 2012, ANVUR and CAE collaborated on two experimental assessments of generic learning outcomes of tertiary students. The instrument, called Test sulle Competenze (TECO), was a translated and adapted to Italy's version of CAE's CLA+, a performance-based assessment that measures generic skills at the tertiary level within the United States and internationally. This choice was made by ANVUR for the following reasons: CLA+ is a ready-to-use standardized assessment; the generic skills are essential higher-order skills that cut across academic disciplines; and the results can be benchmarked against graduating university students in the United States. The TECO was administered to students graduating from certain Italian universities that decided to collaborate on the project.

Since ANVUR aimed at an indirect measure of the program quality, only students who were qualified to graduate as defined by their academic progress were eligible to participate. Accordingly, the eligible students for TECO 2013 were those enrolled in the third or fourth year of a three-year course, or single-cycle master's course, who had acquired all the requisite “basic and characterizing” study credits. Instead, the eligible students for TECO 2015 were those enrolled in the third year (three years earlier) of a three-year course or single-cycle master's course who had acquired 75% of the requisite “basic and characterizing” study credits. In 2013, 5,853 students from 12 participating Italian institutions completed a TECO. Following the 2013 administration, in 2015, a new cohort of 6,268 students from 23 institutions participated in this second study.

In addition to the low turnout of students who were asked to take the test (less than 30%), the other shortcomings of these experiments were the non-representative sample of students, the adaptability of the methodological system proposed by CAE, and the high cost of the whole procedure. Both TECOs were research exercises that depended on voluntary participation of students; thus, when analyzing the data, it is difficult to adequately identify and correct for the self-selection bias. All these issues led to the redefinition of the entire methodological framework of the project.

The new TECO (TECON) will be delivered to all students enrolled in the third year of their undergraduate education, because this is the most numerous, least self-selected sample that is potentially about to enter the labor market. The measurement will be done twice; once at the beginning of undergraduate coursework (or at the end of secondary school), and once just before the end of their course of study. The test results will be certified to students that request them. However, aggregated results at the program level achieved by all participating institutions will be returned to the program's coordinators and institutional governing bodies to be used as a tool for self-evaluation procedures. In any case, the test will not count as an exam and its result will not contribute toward the degree final mark.

While the disciplinary skills test (TECOD) will involve the classe di laurea1 level and be coordinated by ANVUR, the generic skill test (TECO) will be produced and coordinated by ANVUR. Both tests will consist of selected-response questions.

---

1 grouping of study programs defined by Italian law on the basis of disciplinary affinity
questions only, which are computer-delivered during sessions organized by institutions within a defined time window. The objective of TECOD is to provide each study program with comparable information about the outcome of disciplinary learning for their students to use for self-assessment procedures. For this purpose, it is necessary, within each classe di laurea, to identify and define which aspects of the discipline should be the subject of measurement and what minimum levels should be achieved by the students.

In conclusion, the collaboration between ANVUR and CAE was an important first step in formally assessing generic skills in higher education in an international context. TECO was the first example in Italy of assessment of generic skills by a public administration, and the first two experiments had stimulated an intense discussion within the Italian higher education institution system. ANVUR believes that the measurement of the learning outcomes is an essential tool to monitor the quality of the learning process experienced by students. For this reason, we aim at keeping the Agency constantly in the loop of international research in this field and possibly at collaborating with other institutions to try to solve the issues arising from such activities.
The Italian experience
TECO 2012/2013* and 2014/2015

Alberto Ciolfi – ANVUR

Critical Policies and Relevant Applications for an International Comparison of Assessment Results
Cambridge, UK - February 2\textsuperscript{nd}, 2017

*Full TECO 2012/13 Report is provided here:
**Introduction – ANVUR’s tasks**

Among its tasks, Law 76/2010 asks ANVUR to assess not only the processes and the inputs of the national higher education system, but also “the quality of the results and products of management, teaching and research activities... (Art. 3, c.1)”... constitute also object of evaluation the learning outcomes...(Art. 3, c. 2)

Subsequently, Law 240/2010 (Art. 5, section 3) and Legislative Decree 19/2012 initiated the process leading to the integrated Self-Assessment, Periodic Assessment and Accreditation system (AVA).

According to AVA procedures, the actual results achieved by University students in terms of both specialist and generic learning outcomes must not only be compared with the expected ones, but also certified. In the future they will be considered for the assessment & accreditation purposes.
The AVA system today

- The AVA process provides to institutions and single programs indicators relative to:
  - Students’ characteristics in entrance (secondary school type, graduation examination result…);
  - students' careers (dropout rates, time for graduation);
  - students’ satisfaction (in development);
  - job opportunities for graduates.

- However none of these surveys provides a direct measure of the outcome that can be compared between institutions and/or between degree programs within the same institution.
Why measuring competences in higher education institutions ...

There are also substantial reasons that lead ANVUR to assess achieved learning outcomes:

- the development of higher-order skills could be helpful for the next generation’s workforce in order to meet the demands of careers evolving in the 21st century;
- assessments that only foster the recall of factual knowledge have been on the decline, whereas assessments that evoke higher-order cognitive skills, such as analytic and quantitative reasoning, problem solving, and written communication are on the rise;
- Various stakeholders (firms, students, teachers, Central Government) are more and more eager to have an external evaluator assessing and certifying the level of achieved competencies of University graduates.
...can be a challenge

Along with issues related to the framework definition and methodological choices (some important aspects cannot be measured by a test), there are others related to the international perspective:

- The comparison across countries presents challenges due to differences in educational systems, socio-economic factors, and perceptions as to which constructs should be assessed.
- While cross-country comparisons are feasible at the secondary school level (PISA by OCSE), tools that allow for cross-cultural comparisons at the tertiary school level are much less common.
Why CLA+?

CAE’s Collegiate Learning Assessment (CLA+) is a performance-based assessment that measures higher-order generic skills at the tertiary level within the United States and internationally:

- (Almost) ready-to-use standardized assessment
- Generic Skills are essential higher-order skills that cut across academic disciplines
- Their results and outcomes could be benchmarked against graduating university students in the United States.
From CLA+ to TECO

- ANVUR selected one CLA+ International test to be translated, adapted to Italy and validated by a sample of Italian students.
- The goal of the translation and adaptation process for these two studies was to localize (‘Italianize’) the assessment to be consistent with the culture, history and context of students’ home country, Italy. This process ensured that TECO was as much as possible analogous and equivalent to the American version. ANVUR pretested the translated and adapted test and conducted a series of cognitive labs.

The final result was rebranded as TECO (*TEst sulle COmpetenze*).
TECO structure

The TECO is identical for all study courses, consists of two modules.

- “Performance Task” (PT)
- “Selected Response Questions” (SRQ)

- Students were given 60 minutes for the PT and 30 minutes for the 20 (25) SRQs
Performance Task (PT)

For the PT, students are given a scenario and asked to make a decision or recommendation after analyzing a document library that contains various sources of information, such as letters, maps, and graphs. They are then expected to write an argumentative response to the scenario where they justify their decision/recommendation and provide reasons and evidence against the opposing argument(s).

Subscales:

- analysis and problem solving: identifying, interpreting, evaluating, and synthesizing pertinent information and proposing a solution in terms of how to proceed in case of uncertainty
- writing effectiveness: producing an organized and cohesive essay with supporting arguments
- writing mechanics: demonstrating command of written native language.
Selected Response Questions (SRQ)

Set of 25 selected-response questions (SRQs) that are also document based and designed to measure the same construct as the Analysis and Problem Solving sub-score of the PT.

Subscales:
- scientific and quantitative reasoning (e.g., making an inference),
- critical reading and evaluation (e.g., identifying assumptions),
- critiquing arguments (e.g., detecting logical fallacies).
Which students to test?

Since ANVUR aimed for an indirect measure of the study programme quality, only students who were qualified to graduate as defined by their academic progress were eligible to participate. Accordingly, the eligible students for TECO 2013 were those enrolled in the 3rd or 4th year of a three-year course or single-cycle master’s course who had acquired all the required “basic and characterizing” study credits. Instead, the eligible students for TECO 2015 were those enrolled in the 3rd year (three years earlier) of a three-year course or single-cycle master’s course who had acquired 75% of the required “basic and characterizing” study credits.
Participants

In 2013, 5,853 students from 12 participating Italian institutions completed a TECO. Following the 2013 administration, in 2015, a new cohort of 6,268 students from 23 institutions participated in this second study.

<table>
<thead>
<tr>
<th>TECO</th>
<th>Eligible Students</th>
<th>Tested Students</th>
<th>Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>21872</td>
<td>5853</td>
<td>26.76%</td>
</tr>
<tr>
<td>2015</td>
<td>29580</td>
<td>6322</td>
<td>21.37%</td>
</tr>
</tbody>
</table>
Two premises before the ITA/USA comparison

- In TECO 2013, ANVUR decided to limit the SRQ section to 20 instead of 25 questions due to concerns with timing and content.
- The committee felt that the students, unfamiliar with the testing format, would find it difficult to complete 25 questions within the 30-minute timeframe.
- Because the students in 2013 did not exhibit behavior indicating they found the testing format difficult, for TECO 2015, the full set of 25 SRQs was selected from CAE’s bank of questions.
Two premises before the ITA/USA comparison

As part of the experimental design for the 2015 study, a group of experts appointed by ANVUR and trained by CAE, produced set of 25 SRQs developed and written in Italian (SRQ ITA).

A one-way ANOVA to compare the mean difficulty level across the eight forms showed a significant difference in difficulty: the ITA items (Mean = 9.69; SD = 3.34) were more difficult than the USA items (Mean = 11.56; SD = 4.19).

As a result of the difference in mean scores across the 8 forms, the SRQs were linearly equated to the USA set, and scaled.
CLA+/TECO Test Level

<table>
<thead>
<tr>
<th>Descriptive statistics</th>
<th>TECO 2013</th>
<th>TECO 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRQ</td>
<td>PT</td>
</tr>
<tr>
<td>Number of Items</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Number of Students</td>
<td>5853</td>
<td>5853</td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Max</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>Mean</td>
<td>12.31</td>
<td>9.17</td>
</tr>
<tr>
<td>Median</td>
<td>13.00</td>
<td>9</td>
</tr>
<tr>
<td>St. Dev.</td>
<td>2.85</td>
<td>2.95</td>
</tr>
</tbody>
</table>
Cross-country Comparison (PT)

- In a cross-country comparison of students’ critical thinking and written communication skills, results show that the Italian (ITA) students’ performance on the CLA+ is roughly comparable to the results attained by their American (USA) counterparts for the PT.

<table>
<thead>
<tr>
<th></th>
<th>Tested students</th>
<th>Mean score</th>
<th>SD</th>
<th>Percentiles scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25&lt;sup&gt;th&lt;/sup&gt;</td>
</tr>
<tr>
<td>TECO</td>
<td>ITA</td>
<td>5853</td>
<td>1000</td>
<td>852</td>
</tr>
<tr>
<td>2013</td>
<td>USA</td>
<td>4380</td>
<td>1067</td>
<td>937</td>
</tr>
<tr>
<td>TECO</td>
<td>ITA</td>
<td>972</td>
<td>1108</td>
<td>978</td>
</tr>
<tr>
<td>2015</td>
<td>USA</td>
<td>516</td>
<td>1102</td>
<td>1008</td>
</tr>
</tbody>
</table>
Cross-country Comparison (SRQ)

- For the critical reading and scientific and quantitative reasoning skills, results varied. In 2013, Italian students outperformed American students, whereas in 2015, the opposite is true. This could be due to the increase in the number of SRQs in 2015 compared to 2013 and the higher difficulty of the SRQ_ITA

<table>
<thead>
<tr>
<th></th>
<th>Tested students</th>
<th>Mean score</th>
<th>SD</th>
<th>Percentiles scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25th</td>
</tr>
<tr>
<td>TECO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>ITA</td>
<td>5853</td>
<td>1000</td>
<td>908</td>
</tr>
<tr>
<td></td>
<td>USA</td>
<td>4380</td>
<td>796</td>
<td>681</td>
</tr>
<tr>
<td>TECO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>ITA</td>
<td>972</td>
<td>1052</td>
<td>931</td>
</tr>
<tr>
<td></td>
<td>USA</td>
<td>509</td>
<td>1126</td>
<td>1000</td>
</tr>
</tbody>
</table>
Other important variables

- All participating students were required to answer additional survey questions including composition of the household, family socio-economic status, off-site or working status, any form of financial support for studying, diploma and University grades, national or local admission test scores, their perception of whether they had acquired competences in their course of study, and attendance regularity.

- All these data are useful to identify the key predictors of a better performance to the test. In general it is possible to obtain a complete picture of top and bottom performing students.
What went bad with TECO

- Low participation of students who were asked to take the test,
- Low correlation of the scores given by different evaluators to the PT,
- The correlations between the SRQ and the PT subscores and total score, although statistically significant ($p < .01$), are unusually low ($r = .29$)

Note that both TECO 2013 and 2015 were research exercises that depended on voluntary participation of students. Thus, it is difficult when analysing the data to adequately identify and correct the self-selection bias.
First conclusions – international view

- The collaboration between ANVUR and CAE was an important first step in formally assessing generic skills in higher education in an international context.
- This study was the first of hopefully many studies measuring student learning gains and outcomes in tertiary education.
First reactions – Italian context

The TECO is the first example in Italy of assessment of generic skills by a public administration.

This has opened up a huge topic of discussion within the higher education institution system. Is it reasonable to use a measure of learning outcomes to compare study programmes and Institutions? Could the TECO results be used by programmes coordinators to better design course contents?

What is Ministry of education going to do with such data?
Relevant complaints (I)

- Higher education institutions teach disciplinary – not generic – skills! ANVUR should not evaluate a study programme quality only by the levels of generic skills achieved by students.

- Answer: combine the generic skill test, TECO, with a disciplinary test, TECOD, for those fields where Italian experts and Universities are willing and able to construct such a test above content, meaning that TECOD should not assess knowledge but basic disciplinary competencies.
What’s TECOD?

In 2014 ANVUR has established a working group for each field* with the aim to define the minimum set of knowledge and skills that a graduating student should have acquired

- three TECOD were tested in 2015 with a 30 min SRQ test following the 90 min TECO

*19 in total: mathematics, physics, earth science, chemistry, biology, medicine, agricultural science, veterinary medicine, architecture, engineering, literature, languages, psychology, pedagogy, history and philosophy, economics, law, political science, social science
A more accurate assessment of learning outcomes requires a measurement of the initial level (value-added approach). Higher education institution should be assessed according to the student’s gain (hopefully) from the initial conditions, rather than on the basis of the absolute levels reached by their students.

Answer: It could be accomplished for both generic and disciplinary skills. The first with a test at the end of the high school, the latter with the admission tests (already existing for many disciplines)
Future directions in Italy

- It is crucial for ANVUR to assess and certify the generic competences acquired by University students.
- The new Ministerial Decree (Dec 2016) defines quantitative indicators that ANVUR must use for external evaluation. For the first time, the Minister “asked to develop” (are in the Decree but are not defined) indicators relative to student’s learning outcomes and opinions.
- ANVUR must also define, calculate and make available to all institutions a broader set of indicators to help the self-evaluation activities.
Test all student’s enrolled in the 3° year (bachelor’s): this is the most numerous, less self-selected sample, potentially about to enter the labor market;

The measurement must be done twice: at the beginning of bachelor’s courses (or at the end of secondary school) and just before the end of programme.

The test result is certified to students that request it. However, aggregated results (at the programme level) achieved by all participating institutions are returned to the programme’s coordinators and institutional governing bodies.

The test does not count as an exam and its result does not contribute to determining the degree final mark.
New action plan 2016-2019 (II)

- Measured learning outcomes will refer in a distinct manner to disciplinary (TECOD) and generic (TECO).
- Both tests will consist in selected-response questions only, computer-delivered during sessions organized by institutions within a defined time window.
- TECO is produced and coordinated by ANVUR (top->down)
- TECOD is defined and produced at the classe di laurea level and coordinated by ANVUR (bottom->up)
Timeline

2016 fall

TECO Pilot Test
Literacy + Numeracy
5 Institutions – 1000 students

2017 spring

TECO Pilot Test
Problem Solving
5 Institutions – 1000 students

2018 spring

TECO
Literacy + Numeracy
+ Problem Solving
20 Institutions?

TECOD
consultations with academia & scientific societies

TECOD
Test delivering for those disciplines that defined a framework and produced a test
Thank you for your attention

alberto.ciolfi@anvur.it
The Mexican Experience Assessing Generic Skills as Part of the AHELO Project
Patricia Rosas, University of Guadalajara

The assessment of students in the Mexican educational system has been based on multiple-choice standardized tests more than on performance tasks, and this was an important reason to participate in the generic skills (GS) assessment as part of the OECD AHELO project.

For the GS strand, we used two instruments: a questionnaire of multiple-choice questions (MCQs), which aimed to evaluate comprehension processes, fluency, problem-solving, and reasoning; as well as a questionnaire based on constructed-responses tests (CRTs), which evaluated analytical reasoning, problem-solving, and written communication.

We considered this last kind of instrument as an opportunity to experiment in the field of performance assessment. That’s why Mexico has applied the “Museum” CRT in addition to the two performance tasks “Catfish” and “Lake to River,” which were assessed by the rest of participant countries.

The study included 13 Mexican Higher Education Institutions: Universidad Autónoma de San Luis Potosí, Universidad de Guadalajara, Universidad Autónoma de Yucatán, Tecnológico de Monterrey (Monterrey Campus), Instituto Politécnico Nacional, Instituto Tecnológico Superior de Irapuato, Universidad Autónoma de Chihuahua, Universidad Veracruzana, Universidad Autónoma de Coahuila, Universidad Autónoma de Zacatecas, Universidad de Ciencias y Artes de Chiapas, Universidad Politécnica de Aguascalientes, and Universidad Tecnológica de la Mixteca. Sampling was randomly stratified with a sample size of 100 students for each of the participating higher education institutions.

The evidence obtained from the tests showed that the MCQ instrument favored science educational programs, while students of social sciences received more deficient scores, which means that the test did not measure generic abilities. In the CRT, students showed similar results regardless of the educational program they attended, which indicates that the performance task measured the generic skills of the students as it focused on their general training and not on the disciplines. Therefore, we found that the instruments of MCQs and CRTs measure different things.
Mexican experience in
AHELO Generic Skills strand

PhD Patricia Rosas
University of Guadalajara, México

February the 2nd & 3rd, 2017, Cambridge, UK
Content

1. Mexican context
2. Implementation
3. Analysis
4. Lessons learnt
5. Challenges of a future ILOA
OCDE make a wide world call for the HE systems to participate in a feasibility study to see if it was possible to compare internationally students learning outcomes.

The study assessed three strands:
- Economics
- Engineering
- Generic Skills
• 17 countries joined the Project: Abu Dhabi, Australia, Belgium-Flanders, Canada (Ontario), Colombia, Egypt, Finland, Italy, Japan, Korea, Kuwait, Mexico, Netherlands, Norway, Russian Federation, Slovak Republic, United States (CT, PA, MO)

• The project lasts from 2008 to 2013

• The results are described in three volume report available at www.oecd.org/edu/ahelo
1. Mexican context

Reasons to participate

• The need to emphasize from qualitative approaches learning assessment to measure the development of student skills

• In a rapid changing world with broad access to information, that demands new abilities to deal with labor market
1. Mexican context

Reasons to participate

• In a country characterized by its regional, cultural & socio-economical diversity, its greater demand for higher education, and its many educative subsystems, we wanted to know if it was possible to compare LO among educative subsystems at a national level and with other international HEIs.
1. Mexican context

Reasons to participate

• Mexican current mechanisms of evaluation focus mainly on inputs & processes, while learning outcomes assessment has been gaining ground internationally.

• AHELO represented a possibility to know more about the impact that typical indicators such as qualifications of faculty, infrastructure, curriculum updating, tutoring, departmental tests, among many others, has in LO.
2. Implementation

• In the Mexican educative system predominates standardized tests. Our students have few experience to deal with performance trials.

• AHELO opened an opportunity window to emphasize new assessment approaches for higher education system. That’s why we decided to engage in the three strands AHELO proposed.
2. Implementation

Geographical distribution of HEIs that joined AHELO
2. Implementation

- 14 Mexican HEIs joined AHELO
  - 13 for GS
  - 10 for EC
  - 10 for ENG

- 9 of 14 Mexican HEIs applied for the three strands

- 8 of 14 were public universities; 2 were technological institutes (1 public, one private); 1 polytechnic institute; 1 polytechnic college; 1 technical university; 1 multicultural university.
2. Implementation

Quick facts from GS AHELO

- 3 RCT
- 1 MCQ
- 1 Context questionnaire
- 1 Motivation survey
- 5 Training workshops for tests administrators
- 2 training workshops to learn how to design PTs

- Students
  - 1842 (AHELO)
  - 2583 (3rd PT)
- Faculty
  - 400
- Coordinators
  - 13
- Test administrators
  - 56 (AHELO)
  - 150 (PT3)
2. Implementation

Evaluated tests in GS

• Like the countries that participated in AHELO, Mexico evaluated for the generic skills:
  • The two CAE’s CRT: Catfish and Lake to River
  • 40 MCQs taken out of the 55 questions designed by ACER, at random.
  • A context questionnaire.

• In addition, a third performance task designed by CAE was applied (PT3, Museum) and a motivation survey accompanied this evaluation.
2. Implementation

Evaluated tests in GS

- The evaluations for GS were assessed according to the cognitive abilities set by CAE: analytical reasoning, problem solving, and written communication.

- The multiple choice evaluation was designed to assess the following abilities: comprehension processes, fluency, problem-solving, and reasoning.
2. Implementation

Sampling

• Due to difficulties to reach the needed response rates, although participant countries had chosen the target population, they agreed to obtain the sample in different ways based on the specific needs or interests of these countries giving the possibility to have no random samples. This situation does not allow to calculate possible errors attributable to the sample
2. Implementation

Sampling

• In Mexico, we had a random stratified sampling with a sample size of 100 students for each of the participant HEI
• We determined the strata by region, educational program, modality, gender, disciplinary fields, and university entrance cycle
• Similarly, in the case of PT3 we applied stratified random sampling and it we agreed that if any HEI wanted to extrapolate conclusions, it could expand its sample
2. Implementation

Response rate of HEIs in GS

<table>
<thead>
<tr>
<th>IES</th>
<th>Scheduled</th>
<th>Asistents</th>
<th>%</th>
<th>Scheduled</th>
<th>Asistents</th>
<th>PT3</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>UASLP</td>
<td>190</td>
<td>153</td>
<td>81</td>
<td>607</td>
<td>452</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>UDG</td>
<td>196</td>
<td>157</td>
<td>80</td>
<td>1458</td>
<td>1107</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>UADY</td>
<td>190</td>
<td>160</td>
<td>84</td>
<td>91</td>
<td>73</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>UACH</td>
<td>194</td>
<td>154</td>
<td>79</td>
<td>94</td>
<td>81</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>UAZ</td>
<td>189</td>
<td>157</td>
<td>83</td>
<td>100</td>
<td>71</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>UADEC</td>
<td>196</td>
<td>63</td>
<td>32</td>
<td>100</td>
<td>30</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>IPN</td>
<td>183</td>
<td>122</td>
<td>67</td>
<td>1429</td>
<td>205</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>ITESM</td>
<td>187</td>
<td>80</td>
<td>43</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>UTM</td>
<td>151</td>
<td>142</td>
<td>94</td>
<td>100</td>
<td>92</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>UV</td>
<td>196</td>
<td>154</td>
<td>79</td>
<td>354</td>
<td>262</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>UPA</td>
<td>200</td>
<td>161</td>
<td>81</td>
<td>79</td>
<td>68</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>ITESI</td>
<td>200</td>
<td>166</td>
<td>83</td>
<td>100</td>
<td>77</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>UNICACH</td>
<td>200</td>
<td>173</td>
<td>87</td>
<td>100</td>
<td>65</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>Totales</td>
<td>2472</td>
<td>1842</td>
<td>75</td>
<td>4612</td>
<td>2583</td>
<td>56</td>
<td></td>
</tr>
</tbody>
</table>
Response rate of HEIs in GS

- In this type of study, the response rate is 75%, and a non-response rate of 25% could be expected. So at a national level we reached the average for the purpose of the international sample.

- For PT3 some institutions did not reach the sample. However, it is possible to do analysis with the rest of HEIs, since the type of sampling (stratified random) allows it.
In the ACER international report, the data presented an average of 500 and a standard deviation of 100, which makes the performance homogeneous.

About 70% of students score between 400 and 600.

Differences of 20 points or less represent “small” effects, differences of 50 points represents “average” results, and differences of 80 points represent “big” effects.
This graph with transformed (a) and untransformed data (b) clearly shows the effect of such transformation. Note that in "b" the MCQ is below the rest.
3. Analysis

Methods used in the analysis of results

• We used descriptive techniques to observe the behavior of the scores

• We applied psychometric analysis per item in MCQs (Rasch model)

• We also calculated the $\alpha$ Crombach Coefficient

• We estimated the empirical density (kernels smoothing), under the assumption of equal distributions
There is no significant correlation between the instruments
3. Analysis

Comparison of CRT - MCQ scores with 95% confidence bands

Instruments measure different things

Equal distribution hypothesis is rejected
A student of Arts and Humanities requires more ability to correctly solve the MCQ34
3. Analysis

It can be concluded that there is no evidence that the instruments measure the same:

- The hypothesis of equal distribution is rejected
- The correlation coefficient is low (0.27)
- $\alpha$ Cronbach Coefficient is also low (0.47)
MCQ tendency

We observed that the MCQ favor career profiles with more content on exact sciences, which can affect the performance of students coming from other educational programs, therefore it seems that their results do not depend only on the generic skills.
3. Analysis

The statistical analysis of CRTs shows similar performance for students of different educational programs; that is to say, it depends on the generic formation and not of the specific one.
3. Analysis

CRT score comparison

The graph shows that the performance of students from different educational programs in CRT from different contexts was not significantly different; that is, the performance in GS test is not influenced by either the profession or specific knowledge strand.
Comparison of the score distribution for the three PTs.
4. Lessons learnt

• The conceptual idea of AHELO was innovative and suggested many challenges about which we have been reflecting and learning

• We knew that this was a feasibility study and that it could result in a viable or nonviable AHELO

• Comparisons could not be given properly since there were no the same RR among HEIs

• The contextual dimension was not exhaustive
4. Lessons learnt

• The evidence about the validity of the instruments (MCQs) may not be contundent as expected

• Since in the Mexican educational system predominates the application of standardized tests, our students have little experience in performance task evaluation. Nevertheless, the results of motivational survey showed that they were pleased to solve tests in which their opinion was requested beyond a closed answer
4. Lessons learnt

- The integration of MCQ was valuable, but there was no time to discuss this instrument between countries that was not integrated into the conceptual framework. Also, the ACER Consortium report added the CRT and MCQ scores as if they came from the same construct.
4. Lessons learnt

- From the analysis we had in the Economics strand, we believe that it is not only possible but also desirable to evaluate specific skills under CRT to better understand how our students think and act as well as our institutional contribution to developing that profile.
5. Challenges of a future ILOA

• It is necessary to have further discussion on what GS should we —and are able- to evaluate if we want to make international comparisons

• Besides that, it would be good to think how could we integrate GS with specific abilities in only one test to save resources.
5. Challenges of a future ILOA

- It is crucial to have sufficient time to discuss, reach consensus and adjust instruments, as we did during the first stage of AHELO.

- We also have to think about some possible crosses of an ILOA with other variables such as college entrance exams, recognized quality programs, job placement rates ...
5. Challenges of a future ILOA

- The international consensus for a LOA pass for the expectations of governments and nations that will vary in function of their own economic, social and cultural contexts.

- The financial crisis is an important constraint in designing public policies for HE, tempting governments to use this type of evaluations as a way to allocate budgets. We must resist using this type of assessments for this purpose.
5. Challenges of a future ILOA

What would an ILOA do better than separated nations?

• Reach consensus about standard generic capacities needed for this global era

• Help to improve assessing national instruments by considering this global perspective
5. Challenges of a future ILOA

What an ILOA would do better than separated nations?

• Provide insights to make more appropriated ranking systems according to the different HEIs profiles

• Contributing to change the culture of high stakes assessing to low stakes, reducing pressure for allocating budget in function of rankings, and creating a new culture of learning for its sake
Gracias!!!

Session 3: CLA+ Research and Development

Research Cases for CLA+

Doris Zahner, CAE
Roger Benjamin, CAE

CLA+ is a valid and reliable instrument (Zahner, 2013) that measures critical-thinking and written-communication skills of university students. CAE measurement scientists have varied research interests; although, at an elevated level, all research pertains to the study, the use, and/or the improvement of CLA+. A highlight of relevant studies pertaining to the use of CLA+ in the international context was presented.

First, research on predictive validity indicates that the instrument is predictive of university outcomes as measured by cumulative senior (graduating) grade point average (Zahner & Steedle, 2012) and post-university outcomes as measured by employment, full-time employment, salary, and admission into a graduate school program (Zahner & James, 2016). The results from the post-university outcomes study was particularly interesting because there is a significant proportion (approximately 20% to 30%) of minority (Asian, Hispanic, African-American) students from less-competitive institutions who show proficiency in the critical-thinking and written-communication skills measured by CLA+. This finding is of particular importance as it can address an accessibility gap between employers who want to diversify their workforce and qualified minority students seeking employment. A follow-up study with actual employers confirms the assertion that critical-thinking and written-communication skills are important in the workplace. Additionally, students with higher CLA+ scores tended to have higher ratings and ranking as evaluated by their managers than those with lower CLA+ scores (Zahner & Lehrfeld, 2017).

Next, a study of the effect of motivation on CLA+ performance showed that higher motivation resulted in better performance and higher reported levels of effort and engagement (Lehrfeld, & Zahner, 2017). This is of importance in the international context because increasing student motivation will yield truer results of student performance. Motivation can be increased by raising the stakes associated with the assessment, such as providing performance-based badges, individual student reports, and the potential to find employment through matching employers with students who demonstrate proficiency in critical-thinking and written-communication skills.

A third study investigated the performance of students attending for-profit institutions compared to demographically similar students from public and private non-profit institutions. Findings show that, despite its negative image in the United States, students at for-profit institutions had similar critical-thinking and written-communication skills as their non-profit counterparts (Lehrfeld, 2016).

Finally, an example of the educational resource allocation issue in the United States was presented. The major issues in higher education include increasing attainment for students from under-represented groups, improving teaching and learning, and providing verified levels of mastery in critical-thinking and written-communication skills for graduating students. However, the historical inflationary growth of the HEPI means we can assume little or no effort has been made to cut costs over time. Additionally, any new resources to higher education are consumed by cost increases. Therefore, there are fewer resources to attack challenges such as the need to raise access, retention, graduation rates, and improve the quality of student learning. The assessment of learning outcomes is central to the evaluation of criteria of academic programs such as quality, centrality, student demand, cost, and comparative advantage.
Research Cases for CLA+

February 2, 2017 | Cambridge, UK
Overview

• Predictive Validity
  – Admissions or placement decisions
  – Hiring or graduate school decisions

• Motivation
  – Attaching stakes to assessments

• For- vs. Non- Profit
  – public vs. private institutions
PREDICTING UNIVERSITY OUTCOMES
ZAHNER & STEEDLE, 2012
Predicting University Outcomes

• Longitudinal Study from Fall 2005 – Spring 2009

• 50 institutions
  – 9,167 students in Fall 2005
  – 3,137 students in Spring 2007
  – 1,330 students in Spring 2009.

• Attrition mostly due to institutions dropping out of the study.
Predicting University Outcomes

Mean percent variance of university GPA accounted for by HSGPA, SAT, and CLA

<table>
<thead>
<tr>
<th>Year</th>
<th>HSGPA</th>
<th>SAT</th>
<th>CLA</th>
<th>HSGPA &amp; SAT</th>
<th>HSGPA &amp; CLA</th>
<th>SAT &amp; CLA</th>
<th>HSGPA, SAT, &amp; CLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>23.7</td>
<td>17.6</td>
<td>12.7</td>
<td>31.2</td>
<td>29.5</td>
<td>22.8</td>
<td>33.5</td>
</tr>
<tr>
<td>4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>20.8</td>
<td>15.4</td>
<td>11.6</td>
<td>28.3</td>
<td>31.5</td>
<td>21.1</td>
<td>31.5</td>
</tr>
</tbody>
</table>
Predicting University Outcomes

- HS GPA best predictor of university GPA
- SAT + HSGPA were the best predictors of second year GPA
  - An additional 2.3% of the variance was accounted for when adding CLA+ to the model
- CLA + HSGPA were the best predictors of fourth year GPA
  - No additional increase in variance when adding SAT to the model
- CLA measures something different than traditional predictors of university success
Predicting University Outcomes

Relevance to CLA+ International

• Generic skills are indeed differentiated from content knowledge

• Measuring generic skills provides additional information about student learning outcomes
Predicting University Outcomes

Relevance to CLA+ International

• What is currently being used in your country to measure student learning outcomes?

• How might CLA+ be used in this context?
POST-UNIVERSITY OUTCOMES
STUDY 1
ZAHNER & JAMES, 2016
Post-University Outcomes, Study 1

- University seniors – CLA+ in spring 2014
- Surveyed 3, 6, and 12 months following graduation
- Approximately 13,000+ students
  - 1,585 agreed to participate in the survey
  - 993 persisted through all three phases
# Post-University Outcomes, Study 1

Predictive Validity of CLA+ and Other Variables on Post-University Outcomes

<table>
<thead>
<tr>
<th>CLA+ and Variables</th>
<th>Post-university outcomes (all)</th>
<th>Salary</th>
<th>Employ</th>
<th>Full-time employ</th>
<th>Grad school</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAA</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Barron’s</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Field of study</td>
<td></td>
<td>*</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
Post-University Outcomes, Study 1

- Asian
- African-American / Black, non-Hispanic
- Hispanic or Latino
- White, non-Hispanic

Legend:
- Non-Competitive Proficient
- Competitive Non-Proficient
- Competitive Proficient
- Non-Competitive Non-Proficient
Post-University Outcomes, Study 1

Projected National CLA+ Performance

Students Attending Selective Institutions

Students Attending Less-Selective Institutions
Post-University Outcomes, Study 1

• CLA+ is predictive of post-university outcomes such as employment and salary.

• Use of CLA+ scores can address some of the challenges employers and recent graduates face by highlighting these critical skills.

• Assessments like the CLA+ serve as both an effective instrument for identifying high-achieving students from less and non-competitive institutions and making their skills more visible to perspective employees.
POST-UNIVERSITY OUTCOMES
STUDY 2
ZAHNER & LEHRFELD, 2017
Post-University Outcomes, Study 2

- University seniors – took CLA+ in spring 2014
- Fall 2015 – Surveyed 75 employers of original cohort of students
Post-University Outcomes, Study 2

On a scale from 1 – 5,

• How important are the following skills to successful performance in the participant’s position:
  – Analysis and Problem Solving
  – Writing Effectiveness
  – Writing Mechanics

• How would you rate the participant on the following skills:

• Overall, where does the participant’s performance rank compared to other recent university graduates in your workplace?
Post-University Outcomes, Study 2

- APS
- WE
- WM

- Very Important
- Important
- Moderately Important
- Of Little Importance
- Unimportant
Post-University Outcomes, Study 2

Ordinal logistic regression models for predicting participants’ post-university performance as measured by their managers

<table>
<thead>
<tr>
<th>CLA+</th>
<th>APS</th>
<th>WE</th>
<th>WM</th>
<th>Relative Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>t-statistic</td>
<td>17.55**</td>
<td>26.00*</td>
<td>23.80**</td>
<td>16.56**</td>
</tr>
</tbody>
</table>
Post-University Outcomes, Study 2

• Employers find critical thinking and written communication skills to be important or very important for entry-level positions in the workforce.

• Students with higher CLA+ plus scores tended to have higher ratings from their employers.

• Findings from this study offer support for the conclusion that critical-thinking and written-communication skills are important in predicting career placement and workplace success (Arum & Roksa, 2014).
Post-University Outcomes

Relevance to CLA+ International

• CLA+ scores can be used as part of employment decisions

• CLA+ can be used to identify minority or less-represented students

• ANVUR study
Post-University Outcomes

Relevance to CLA+ International

• How can CLA+ scores be used in the post-university context?

• What does the ethnic differentiation look like in your country?
MOTIVATION
LEHRFELD, 2017
United States Army Cadet Command

Performance of ROTC students vs. non-army controls

• Matching pool: 18,276 ROTC students and 34,381 control students from 2014 and 2016
  – 2014 = low stakes year, 2016 = high stakes year

• 1-to-1 matching on propensity scores estimated via logistic regression
  – Student-level matching variables: age, gender, race/ethnicity, parental education, field of study, English as a first language
  – Institution-level matching variables: Carnegie classification, Barron’s selectivity, public vs. private, historically Black institution, institution size

• Matched data sets had in each group (ROTC and control):
  – 4,473 students (2014)
  – 4,420 students (2016)
United States Army Cadet Command

High Stakes
- Total Score
- PT Score
- SRQ Score

Control
- Total Score: 1100
- PT Score: 850
- SRQ Score: 1150

ROTC
- Total Score: 1150
- PT Score: 900
- SRQ Score: 1120
United States Army Cadet Command

Direct comparison of low and high stakes conditions

- Matching pool: 11,545 ROTC students total from 2014 and 2016
- Same matching model as previously described
- Matched data set had 4,420 ROTC students from each year (stakes condition)
United States Army Cadet Command

Effect of Stakes

- Total Score
- PT Score
- SRQ Score

Low Stakes vs. High Stakes

CAE
United States Army Cadet Command

Effect of Stakes

- Engagement
- Effort

Low Stakes

High Stakes
United States Army Cadet Command

- Higher motivation resulted in better performance for ROTC students

- Higher motivation also yielded higher reported levels of effort and engagement

- The effects of motivation persisted even after controlling for many variables related to academic achievement
United States Army Cadet Command

Relevance to CLA+ International

• Increasing stakes with badging increases motivation

• Providing individual student results

• Matching graduating students with potential employers
United States Army Cadet Command

Relevance to CLA+ International

• What are some ways to increase stakes?

• What is the current model for students seeking employment opportunities?

• Is there a centralized agency that can help with matching students and employers?
FOR-PROFIT VS. NON-PROFIT
LEHRFELD, 2016
For-Profit Institutions

Performance of students at private for-profit institutions vs. students at public and private non-profit institutions

• Matching pool: 663 for-profit university students and 24,702 control students from 2015

• 1-to-1 matching on propensity scores estimated via logistic regression
  – Matching variables: gender, race/ethnicity, parental education, field of study (groupings of university majors), and English as a first language

• Matched data set had 624 students in each group
  – Here we report data on seniors only (328 in each group)
For-Profit Institutions

- Total Score
- PT Score
- SRQ Score

For-Profit
- Total Score: 1100
- PT Score: 1050
- SRQ Score: 1100

Non-Profit
- Total Score: 1100
- PT Score: 1050
- SRQ Score: 1100
For-Profit Institutions

- Engagement
- Effort

For-Profit

Non-Profit
For-Profit Institutions

• Despite its negative image in the United States, students at for-profit institutions had similar critical thinking and written communication skills as socio-demographically similar students
  – For-profit seniors did better than matched non-profit seniors on the PT

• Effort and engagement were higher for proprietary students, perhaps reflecting a higher degree of motivation on their part

• More research is needed – particularly in collecting institutional characteristics – to make these findings more robust
For-Profit Institutions

Relevance to CLA+ International

• Private universities may have similar questions pertaining to the skills and abilities of their students

• Results can be used for informational or even marketing purposes
For-Profit Institutions

Relevance to CLA+ International

• What are some of the issues between public and private universities internationally?

• How can CLA+ results be used for private universities and universities in general?
FUTURE RESEARCH
Possible Future Research

• Predictive validity

• Efficacy of programs

• Construct equivalence across countries

• The effect of order on PT and SRQ performance
AHELO PERFORMANCE TASKS
ZAHNER & STEEDLE, 2014
### International CLA PT Comparison

Unadjusted mean PT (Lake-to-River) scores by country

<table>
<thead>
<tr>
<th>Country</th>
<th>Mean PT Raw Score</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korea</td>
<td>23.9</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Finland</td>
<td>19.6</td>
<td>.7991</td>
</tr>
<tr>
<td>Norway</td>
<td>19.5</td>
<td>.9712</td>
</tr>
<tr>
<td>USA</td>
<td>19.5</td>
<td>*</td>
</tr>
<tr>
<td>Colombia</td>
<td>18.3</td>
<td>.0002</td>
</tr>
<tr>
<td>Egypt</td>
<td>17.4</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Kuwait</td>
<td>17.4</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Mexico</td>
<td>16.7</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Slovak Republic</td>
<td>15.9</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
### International CLA PT Comparison

Unadjusted mean PT (Catfish) scores by country

<table>
<thead>
<tr>
<th>Country</th>
<th>Mean PT Raw Score</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korea</td>
<td>22.5</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>USA</td>
<td>18.9</td>
<td>*</td>
</tr>
<tr>
<td>Kuwait</td>
<td>18.6</td>
<td>.5899</td>
</tr>
<tr>
<td>Colombia</td>
<td>18.5</td>
<td>.2478</td>
</tr>
<tr>
<td>Norway</td>
<td>17.9</td>
<td>.1606</td>
</tr>
<tr>
<td>Egypt</td>
<td>17.6</td>
<td>.0002</td>
</tr>
<tr>
<td>Finland</td>
<td>17.6</td>
<td>.0087</td>
</tr>
<tr>
<td>Mexico</td>
<td>17.6</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Slovak Republic</td>
<td>16.0</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
ANOTHER EXAMPLE OF EVIDENCE-BASED DECISION-MAKING: EDUCATION RESOURCE ALLOCATION
BENJAMIN, 2017
Education Resource Allocation
Example from U. S. case

• Major higher education issues include,
  – Increasing attainment---access, retention, graduation, and gainful employment of students from under-represented groups
  – Improving teaching and learning quality
  – Provision of certified, mastery level badges for graduating seniors useful for pursuit of employment
  – Education Resource Allocation, the example here.
Implications of Figure 1

• The historical inflationary growth of the HEPI means we can assume little or no costs have been made to cut costs over time.

• Any new resources to higher education are eaten up by cost increases,
  – Fewer resources to attack challenges such as the need to raise access, retention, graduation rates and improve the quality of student learning available
UIR is the Path Forward

• Use data analytics and CLA+ test results education leaders can use to set academic priorities and cut costs by
  – 1) more efficient allocation of resources to academic departments to reduce costs and raise the quality of student learning outcomes
  – 2) Out sourcing or elimination of non academic support functions through use of new education technology
Assessment of Learning Outcomes

• Is central to the evaluation criteria of academic programs which in the U.S. are,
  – Quality (of learning outcomes)
  – Centrality
  – Student demand
  – Cost
  – Comparative advantage
Examples of Reductions in Costs Assisted by Ed Tech Innovations?

- The library
- The Student Enrollment office
- Competency based model---courses on line
- Flip the classroom---instructor becomes “coach” and content streamed from devices not textbooks
Session 4: CLA+ Collaborators

Adapting and Validating the Collegiate Learning Assessment to Measure Generic Academic Skills of Students in Germany
Olga Troitschanskaia

Associated with the Modeling and Measuring Competencies in Higher Education (KoKoHs2) research program, the German research team in collaboration with CAE adapted and validated CLA+ as the first German-language instrument for assessing university students’ and graduates’ generic academic skills. The objective was also to maximize international compatibility and comparability of the adapted assessment and results. The underlying concept and assessment framework were aligned with those in other countries implementing CLA+, including the United States, Italy, and the United Kingdom.

Milestones
1. translating the U.S. test instrument into German and adapting it to the German culture to obtain a localized German instrument
2. validating the German instrument comprehensively for use in higher education in Germany
3. based on the validation results, exploring the need for further development and adaptation

Project Steps
1. Translation and Adaptation in cooperation with cApStAn and CAE based on the Test Adaptation Guidelines (International Test Commission, 2005, 2016) and TRAPD process (Cross-Cultural Survey Guidelines, Mohler et al., 2016) aiming for functional equivalence between the German and U.S. versions.
2. Validation based on the Standards of Educational and Psychological Testing (AERA, APA & NCME, 2014); including
   - Content validation through curricular analyses, expert panels, and lecturers’ online ratings
   - Cognitive validation using cognitive interviews to analyze students’ response processes
3. Additional cognitive testing of adaptation to ensure comparability is maintained

Future Opportunities
1. field testing and administration in higher education institutions in Germany
2. in collaboration with CAE and further international partners, exploration of the feasibility of international comparative analyses

German project team Prof. Dr. Olga Zlatkin-Troitschanskaia, Prof. Dr. Hans Anand Pant, Miriam Toepper, Dimitri Molerov, Dr. Ramona Buske, and Dr. Sebastian Brückner
Kooperation partner: Prof. Dr. Silvia Hansen-Schirra, Dr. Sascha Hofmann, and Beatrice Danieli

2 In Germany, the Federal Ministry of Education and Research has established a national research program on “Modeling and Measuring Competencies in Higher Education” (KoKoHs). The first funding phase (2011–2015) involved 24 collaborative projects comprising approximately 70 individual projects conducted by almost 220 researchers, focusing on modeling and measuring domain-specific and generic competencies in higher education. The outcomes of KoKoHs research initiative, which also gave the basis for this study, included 40 competency models and more than 100 measuring instruments. The assessments were administered to more than 50,000 students at more than 220 higher education institutions throughout Germany to gather evidence of their psychometric quality (see Zlatkin-Troitschanskaia et al., 2016a; b). In the current funding phase, which runs between 2016 and 2020, the new KoKoHs program focuses on “Validations and Methodological Innovations.”
CLA+ International Meeting
Adapting and Validating the Collegiate Learning Assessment (CLA+) to Measure Generic Academic Skills of Students in Germany

Implications for International Assessment Studies in Higher Education

Olga Troitschanskaia
Agenda

1. Background
2. Aims and Framework of the German Adaptation and Validation Study
3. Project Overview
4. Preliminary Results
   4.1 Expert panels, Curricular analysis, SRQ rating
   4.2 Cognitive Interviews
5. Conclusion and Outlook
1. Background

Globalization, digitalization, demographic change and growing mobility…

→ growing need for valid performance-based assessments of generic higher-order skills that can be used with different groups of students from different countries

Meta-analysis (Zlatkin-Troitschanskaia et al., 2016) shows that there are no German-language instruments for measuring university students’ generic higher-order skills

A German research team from the KoKoHs program (Modeling and Measuring Competencies in Higher Education) collaborated with the CAE to adapt and validate the CLA+ for the German language and cultural context to measure such skills in higher education students and graduates in Germany
2. Aims and Framework of the German Adaptation and Validation Study

**Goals**

enable the assessment of generic higher-order skills in Germany

**How?**

by adapting and validating the CLA+ for a German context while also aiming to ensure the international compatibility and comparability of the adapted assessments and results

*Four major milestones to achieve these goals*

- translating the U.S. test instrument into German and adapting it to the German culture
- validating the German instrument comprehensively for use in higher education in Germany
- exploring the need for further development and adaptation
- conducting international comparability analyses (with the CAE team, partners in Italy & UK)
2. Aims and Framework of the German Adaptation and Validation Study

Concept

No standardised definition of the term “generic higher-order skills”

Common conceptual framework in literature:

Wheeler and Haertel (1993) determined two contexts in which these skills are employed

a) situations where thought processes are needed for solving problems and making decisions in everyday life
b) contexts where mental processes can be applied that have to be developed by formal instruction, including processes such as comparing, evaluating and justifying

CLA+ aims to measure specific aspects

It holistically assesses analytical reasoning and problem solving

With different case-based task formats and both complex performance tasks (PT’s) and selected-response questions (SRQ’s), administered on a computer
## 2. Aims and Framework of the German Adaptation and Validation Study

### Overview of the German study

<table>
<thead>
<tr>
<th>Year/Month</th>
<th>Activity Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 2015</td>
<td>Selection of tasks (PT 1 and 25 SRQs) for the German study</td>
</tr>
<tr>
<td>Summer 2015</td>
<td>Workshop with CAE’s developers of the CLA+, including scorer training</td>
</tr>
<tr>
<td>Summer 2015</td>
<td>Meeting with colleagues from Italian National Agency for the Evaluation of the</td>
</tr>
<tr>
<td></td>
<td>University and Research Systems (ANVUR)</td>
</tr>
<tr>
<td>Summer/Autumn 2015</td>
<td>Agreement on translation guidelines between CAE, German team and translation</td>
</tr>
<tr>
<td></td>
<td>agency cApStAn</td>
</tr>
<tr>
<td>Autumn 2015</td>
<td>Translation by cApStAn (PT 1, 25 SRQs, test instructions, scoring guidelines)</td>
</tr>
<tr>
<td>Autumn 2015</td>
<td>Review and revisions of translation by German team and first adaptation round for PT1</td>
</tr>
<tr>
<td>Autumn 2015</td>
<td>Curricular analyses</td>
</tr>
<tr>
<td>Winter 2015/16</td>
<td>Expert workshop I: group discussion with 10 national experts from different fields of</td>
</tr>
<tr>
<td></td>
<td>studies</td>
</tr>
<tr>
<td>Winter 2015/16</td>
<td>Second adaptation round by German team for PT 1</td>
</tr>
<tr>
<td>Winter 2015/16</td>
<td>Expert workshop II: group discussion with 10 national experts from different fields of</td>
</tr>
<tr>
<td></td>
<td>studies</td>
</tr>
<tr>
<td>Winter 2015/16</td>
<td>Third adaptation round by German team for PT 1</td>
</tr>
<tr>
<td>Winter 2015/16</td>
<td>Translation of PT 2 by cApStAn</td>
</tr>
<tr>
<td>Winter 2015/16</td>
<td>Review and revisions of translation by German team and first adaptation round for PT2</td>
</tr>
<tr>
<td>Winter 2015/16</td>
<td>Expert workshop III: group discussions with 3 translation experts</td>
</tr>
<tr>
<td>Spring 2016</td>
<td>Second adaptation round for PT 2 by German team</td>
</tr>
</tbody>
</table>
## Overview of the German study

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 2016</td>
<td>Meeting with colleagues from UK’s Learning Gain Program (representatives from the Centre for Excellence in Learning and Teaching (CELT))</td>
</tr>
<tr>
<td>Spring/Summer 2016</td>
<td>10 cognitive interviews with students (PT 1 and 2)</td>
</tr>
<tr>
<td>Spring/Summer 2016</td>
<td>Localization of PT 2 by the German team</td>
</tr>
<tr>
<td>Autumn 2016</td>
<td>Back translation of localized PT 2 and review by CAE</td>
</tr>
<tr>
<td>Winter 2016</td>
<td>Online rating by 12 lectures (25 SRQs)</td>
</tr>
<tr>
<td>Winter 2015/ Spring 2017</td>
<td>10 cognitive interviews with students with localized PT 2</td>
</tr>
<tr>
<td>Spring/Summer 2017</td>
<td>Further analyses and documentation of results</td>
</tr>
<tr>
<td>Summer/Autumn 2017</td>
<td>Comparison of the original version from the U.S. and the adapted test versions from the UK, Germany, and Italy</td>
</tr>
<tr>
<td></td>
<td>Exchange of data and further cross-cultural comparative analyses</td>
</tr>
</tbody>
</table>
3. Project Overview

Translation and Adaptation

The TRAPD* process ensures that the test is reviewed, revised and appraised by a variety of experts on its content, methodology, and translation.

Two performance tasks (PT) of the original CLA+ were adapted:

- adaptation for PT 1 involved no major difficulties,
- adaptation for PT 2 explicated cultural differences and therefore difficulties in the translation process were observed.

To maintain the purpose of the CLA+ in the process of translation and adaptation, two techniques were used:

- Ask-the-Same-Question approach: aims to alter the original item composition as little as possible in order to preserve psychometric properties.
- Ask-a-Different-Question approach: aims to translate the original item into a local context.

*Translation, Review, Adjudication, Pre-testing, and Documentation (for adapting international assessments and surveys)
3. Project Overview

Translation and Adaptation

Translation, adaptation and validation process for the 1st PT in the German study
4. Preliminary Results

How first results were obtained...

Conducting three expert workshops (December 2015 – February 2016) in Mainz and Berlin

Experts from different scientific backgrounds*

discussing about...

- construct definition and operationalization
- domain-specificity and generality
- psychometrics
- challenges of trans-disciplinarity
- necessary additions and optional modifications
- scoring problems
- item quality
- relevance to the curriculum for higher education in Germany
- content validity
- suitability

*including lecturers in biology, business and economics, chemistry, English linguistics and translation studies, higher education research, mathematics, medicine, physics, psychology, social and political sciences, linguistics and translation studies from different German universities
4. Preliminary Results

First results

1. Construct definition
   defined more clearly for Germany and linked to theory and empirical data

2. Further development of the adapted test instrument
   assessing and controlling for additional individual student characteristics and influence factors in subsequent validation analyses

3. Relevance to every-day life and sensitivity to study domain
   students from certain disciplines in Germany (e.g. degree courses without an empirical focus) could have a disadvantage in the test

4. Cultural and linguistic comparability
   challenges of linguistic and cultural comparability were identified in particular for the PT’s and their scoring

5. Equivalence of different performance tasks and scoring
   experts compared underlying construct definitions and relations to domains and culture
First results

Expert panels indicated

• CLA+ is an innovative approach to performance assessment

• the assessment format is an interesting and useful addition to current examination practice in Germany

• there are various challenges to be addressed before the instrument can be used in Germany as well as in an international study
4.1 Expert panels, Curricular analysis, SRQ rating

Curricular Analysis

curricula and module descriptions from 32 different degree courses were analyzed

curricular analyses suggested that the adapted item content of the CLA+ is part of curricula in higher education in Germany

in addition, curricular relevance and content validity was supported by the experts, which indicated that the types of skills assessed are being taught at higher education institutions in Germany
4.1 Expert panels, Curricular analysis, SRQ rating

SRQ Rating

12 professors and lecturers from higher education institutions in Germany rated: curricular relevance and difficulty of the items*

Content validity was confirmed for all adapted SRQ items from the CLA+

Constructs of generic skills were understood in a similar way in different study domains at various universities

* for example: “Does the item represent a higher education curriculum or a higher education domain”? 
4.2 Cognitive Interviews

Validation of the adapted PT1 and PT2

How?

cognitive interviews were conducted with 10 students* to explore their understanding of the items as well as to identify and analyze mental processes occurring during the response process.

Why cognitive interviews?

• They assess more complex aspects of the mental processes during task-solving

• A cognitive interview study preceded the field application as a pretest

• Aiming to create functionally equivalent tasks for multiple languages in which CLA+ is used

* sample included beginner and advanced students, students from different study domains and from different performance levels
4.2 Cognitive Interviews

Validation of the adapted PT1 and PT2

Conduction of the interviews

Concurrent phase:
• students worked on the tasks autonomously and without interacting
• the interviewer took notes

Retrospective phase:
standardized interview guideline to ask ten questions on different aspects of the tasks and the solving process*

* for example: “What did you find especially difficult about the task?”
4.2 Cognitive Interviews

Validation of the adapted PT1 and PT2

Preliminary results

time of item responding varies between students

a large part of solving-time was spent on studying the provided documents

students perceived the topics of the tasks as interesting, but were not necessarily motivated to process and solve them

problem situation described in both tasks was perceived as realistic

relations to everyday real life became evident

students would have liked different, more specific contents to prepare them for their professional life
the performance assessment enables measuring higher-order skills at the academic level in higher education

assessments in higher education practice should be accompanied by corresponding teaching-and-learning tools

this kind of assessment is innovative for higher education practice in Germany and has significant potential for enhancing curricula and instruction to promote students’ interdisciplinary skills

further research and development is needed in particular with a focus on the concept and test definition

further examination of domain-specificity is needed
5. Conclusion and Outlook

Further steps

to what extent this assessment is suitable for intra- or cross-institutional comparisons remains to be explored in further research

different adapted versions shall be examined with regard to their measurement equivalence in order to ensure that the adapted tasks measure the same skills and to determine which further adaptations are necessary

eye-tracking studies in order to control the effects of general reading abilities could benefit further research progress
Implications for International Assessment Studies in Higher Education

- Results from OECD’s AHELO have indicated that international comparative assessments are possible, but very challenging in terms of:
  - conceptualization,
  - methodology, and
  - harmonization
- To ensure comparability, administration and interpretations need to be comparable, as well, which requires implementation of similar processes across nations for:
  - pretesting,
  - sampling,
  - task administration,
  - technical presentation,
  - incentivization,
  - validity and
  - comparability analyses
Implications for International Assessment Studies in Higher Education

- psychometric quality criteria need to be confirmed in each country individually based on a comprehensive, validity concept
  - that aligns theoretical and empirical evidence from the test scores and the interpretations of these test scores to indicate
  - whether the test score sufficiently represents the targeted construct

- comparable processes and quality benchmarks need to be
  - linked in validity arguments within countries and
  - with the same quality indicators across countries to maximize both local applicability and international comparability.
Implications for International Assessment
Studies in Higher Education

- the most promising approach: integrate adaptation and test development, aiming for top-down comparability and harmonization at every step

- generating tasks
  - with similar structures and appearances
  - that are adaptable across nations,
  - as well as for subsequent translation and adaptation,
  - and quality assurance in revisions
Implications for International Assessment Studies in Higher Education

• performance assessment involves complex, lifelike tasks and parallel real-life responses that can be
  • intricate,
  • lengthy and
  • limited in number due to time and cost

• standard psychometric models were developed for multiple-choice assessments, with many discrete test items
  • that are scored dichotomously and
  • that are organized into tests designed to measure one clearly defined construct
Implications for International Assessment Studies in Higher Education

- performance assessment is complex to model and implement, and requires thorough testing and examination to confidently associate test-taker performance with a score or a performance category
- some of the challenges associated with modeling performance assessment are
  - limited number of observations
  - complex and varied score scales
  - human influence (raters)
  - human influence (group members)
  - person x task subject-matter interaction
  - connectedness
  - dimensionality
Implications for International Assessment
Studies in Higher Education

- Careful planning and problem resolution routines are needed to address arising questions, such as what to do
  - if items are not adaptable,
  - if target samples sizes are not reached,
  - if administration and incentivization are implemented differently,
  - if quality indicator benchmarks are not met in individual countries
  - etc.
- additional variables need to be assessed and harmonized across participating countries, to
  - enable controlling for educational input and
  - process factors that correlate with good test results
Implications for International Assessment Studies in Higher Education

Theory of learning and knowing

Focus

Assessment

Influence

Curriculum

Instruction

(Pellegrino et al. 2001)
THANK YOU!
Any further Questions?


References


References


Purpose of the PAL Project

Constructive alignment triad

**Intended learning outcomes:**
Which abilities and skills should students achieve by the end of the course or degree course?

**Assessment tasks:**
By means of which examination methods can the intended learning outcomes be assessed in an adequate manner?

**Teaching/learning activities:**
By means of which teaching and learning activities can the intended learning outcomes be achieved?

(based on Biggs & Tang 2011)
Using CLA+ and the UK Engagement Survey to Measure Learning Gain
Stuart Brand and Jamie Morris, Birmingham City University

In this presentation, we described the lived experience in England of a Higher Education Funding Council for England (Hefce) funded project on “Measuring Learning Gain.” This project is one of 13 across England; it involves the use of CLA+ testing. The project is hosted by Birmingham City University (BCU), a post-1992 university, and involves work with three partner universities, all also post-1992 institutions. They are: Coventry University, Liverpool John Moores University, and Staffordshire University.

Our aim from the outset in 2015 was to carry out a longitudinal study following 1,000 students across the four partner universities through the three years of their programs of study. During this time we planned to administer the CLA+ test on four occasions and also at the later testing points to administer some questions from the United Kingdom Engagement Survey (UKES). UKES is significantly derived from the National Survey of Student Engagement (NSSE) in the United States. It has been suggested that it would be interesting to look for correlations between UKES and CLA+ outcomes.

In our presentation, we asserted that perhaps these aspirations were a little naïve. This was because we encountered significant problems in recruiting students to participate in the project. It should be stressed that this challenge of student recruitment is not principally a feature of the use of CLA+, it being reproducible across all projects that asked students to engage in any extra activity beyond their program of study.

We set out in our presentation some successes and challenges that we have encountered. Successes included: recruitment of 845 students in 2015-16 (against an initial target of ~1,400); a positive student response to the test from those who elected to participate; excellent support from CAE in New York; the relative ease of task adaptation; and improvements to student participation where key faculty (academic staff) members in the program team engaged in student recruitment. Challenges included: the question as to whether we can secure ongoing participation from those who participated once; a persuasion task with faculty concerning the purposes of our project; the ownership of outcomes; and the relative lack of student enthusiasm for selected-response questions contrasted with the performance task.

We also described next steps for our project, which include the provision of a second longitudinal study commencing Autumn 2016 and a cross-sectional study in each institution during 2017.
Using CLA+ and the UK Engagement Survey to Measure Learning Gain

Professor Stuart Brand, Jamie Morris
Birmingham City University
Cambridge - February 2017
Birmingham City University context

- Post 1992 University
- ~24,000 Students
- 8 Campuses moving to 2/3
- Vocational approach
- Multicultural student population
- Commuter students (70%+)
The Higher Education Funding Council for England (HEFCE) and the Learning Gain Initiative

- BCU + three partners: Coventry, Liverpool John Moores, Staffordshire (all post-92)
- Use of Collegiate Learning Assessment (CLA+)
- Use Of UKES
- Original proposal: 1,000 students – longitudinal study, 2015-18
- Now revised!
Aspirations – were we naïve?!

• Variety of disciplines across four partners

• Schedule of 4 testing points for CLA+:
  • Early Year One
  • End of Year One
  • End of Year Two
  • End of programme

• BUT?!
In 2002, CAE (Council for Aid to Education) a not for profit based in NYC, launched a national effort to assess the quality of undergraduate education

- CLA+ (Collegiate Learning Assessment) is a web administered, value added assessment programme

- Performance Task and Selected Response Questions
The Collegiate Learning Assessment (CLA+) is a performance based test that measures critical thinking, problem solving, analytic reasoning, and writing skills.

The CLA+ uses a Performance Task and a series of selected-response questions measure these higher order skills.

No prior knowledge of any particular field is necessary in order to perform well.

The principal goal of the CLA+ is to provide an objective assessment about the critical-thinking skills a student possesses as they enter and exit college.

CLA+ post 2011
CLA+ - briefing students

• This is a low stakes assessment – your performance in these tests does **not** contribute to your assessment results on your programme of study.

• However, participation in the CLA+ can help you assess areas of strength in the set of skills that are in demand from employers.

• We can work with you to support improvement in key areas for your skills development.
First iteration (Nov ‘15 – Feb ‘16) and follow up

- The challenge of student recruitment
- Administering test sessions
- Student reaction – is there a follow up?
Successes and Challenges

- Recruitment of 845 students in 15-16
- Interest at time of testing
- Positive reaction to performance task
- The support from CAE
- Task adaptation
- Engaged programme staff = high student participation rate

- Scorer training – timing and ongoing challenges
- Can we keep the students hooked? Which incentives will really work?
- How do faculty (academic staff) feel about the initiative?
- Who does the whole thing belong to?
- The Selected Response Questions issue
Next steps

- Reschedule
- Employ students?
- Focus groups
- Increase faculty buy-in
J. Enrique Froemel from Chile provided an overview of his work recruiting participation in CLA+ International in South America. The focus on Chile, Colombia, Paraguay, and Peru is based on their relatively autonomous education systems, assessment culture, and stated objectives for educational improvement and reduction of student remediation. The initial focus will be on recruiting private universities, with the intention of building a consortium based around the assessment community. At the initial stages of recruitment, ministry level contacts have not been considered, rather prioritizing partnerships with individual institutions and education service companies. Dr. Froemel's main takeaway was that, although there will be problems, they can be overcome and the project will move forward.
PARTICIPATION OF UNIVERSITIES FROM FOUR SELECTED SOUTH AMERICAN COUNTRIES IN CLA+

J. Enrique Froemel, Ph.D.
A NON-TRADITIONAL APPROACH TO THE IMPLEMENTATION OF A COMPONENT OF THE CLA+ STUDY

• Roger and I, were both speakers at the Learning Evaluation Systems in Higher Education International Seminar, held in Cali, Colombia, in 2016

• There, we discussed the CLA+ Project and he asked my help on incorporating South American universities into the project. We concluded that the study was in that respect, more akin to private institutions

• The idea was to develop leadership for recruiting institutions in a sub-region where there are few university associations adequate for the purpose and where public authority action is severely hampered by bureaucracy

• Consequently, the mode of operation was to approach individual private universities
WHY THESE FOUR SOUTH AMERICAN COUNTRIES?

- On a first stage, for various reasons, four countries were suggested. i.e.: Chile, Colombia, Paraguay and Peru
  - All of them have rather open higher education systems, with predominant private components
  - They are presently also at various stages of higher education development, but also all strive for student achievement improvement
  - In most of them, due to lower than acceptable school education quality, higher education students are in dire need of assessment and remediation
  - All of them have developed an assessment culture, through the implementation of national achievement assessment systems and participation in international studies, as well, i.e.: LLECE, PISA, TIMSS, PIRLS, IALS, ALL and LAMP.
BRIEF NUMERICAL DESCRIPTION OF UNIVERSITY SYSTEMS IN THE FOUR SELECTED COUNTRIES

<table>
<thead>
<tr>
<th>COUNTRIES</th>
<th>Nº OF INSTITUTIONS (STUDENTS)</th>
<th>Nº OF PRIVATE (STUDENTS)</th>
<th>Nº OF PUBLIC (STUDENTS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHILE</td>
<td>59 (655,738)</td>
<td>43 (481,496)</td>
<td>16 (174, 242)</td>
</tr>
<tr>
<td>COLOMBIA</td>
<td>82 (1,431,983)</td>
<td>50</td>
<td>32</td>
</tr>
<tr>
<td>PARAGUAY</td>
<td>54 (211,857)</td>
<td>47</td>
<td>7</td>
</tr>
<tr>
<td>PERU</td>
<td>142 (1,107,424)</td>
<td>91</td>
<td>51</td>
</tr>
<tr>
<td>TOTAL</td>
<td>337 (3,407,002)</td>
<td>231</td>
<td>106</td>
</tr>
</tbody>
</table>

UNIVERSITIES FROM FOUR SELECTED SOUTH AMERICAN COUNTRIES
(Data ranges from 2013 to 2016, at official sources)
OTHER INFORMATION ON UNIVERSITY SYSTEMS IN THE FOUR SELECTED COUNTRIES

• Chile:
  • Approximately 90% of students are undergraduate and after 4 to 5 years they obtain the licenciate degree. Doctoral and Master´s programs do exist.
  • A mixed (public autonomous and private) accreditation is in place since 2006. Institutional and some program accreditation is voluntary, although compulsory; for programs in Medicine and Pedagogy it is mandatory. Almost all universities apply for accreditation and approximately 80% get it. Two Chilean institutions are US accredited.
  • Funding to higher education comes from a blend of public and private sources.
  • A reform process for Higher Education is under-way in Congress.
OTHER INFORMATION ON UNIVERSITY SYSTEMS IN THE FOUR SELECTED COUNTRIES

• Colombia:
  • A public autonomous accreditation system has existed for almost two decades. The involved agency also tests students when entering and leaving higher education
  • A long term tradition of quality assurance in education exists, through a national testing and research agency (ICFES)
  • Funding to higher education comes primarily from public sources
• Paraguay:
  • Universities can only created by law, although its number has significantly increased in recent years
  • There is a state based accreditation system, created in 2003, although extremely basic in its breadth of application and quality of processes
  • Most programs lead to undergraduate degrees
  • The system serves mainly Paraguayans, although demand from Argentines and Brazilians is very high in bordering areas
  • There are serious concerns on quality of institutions and programs
  • Higher education is almost totally funded by the State.
OTHER INFORMATION ON UNIVERSITY SYSTEMS IN THE FOUR SELECTED COUNTRIES

• Peru:
  • Higher education has experienced a large boost in recent years
  • An effective accreditation system does exist and a new voluntary program accreditation model was implemented in 2016. Two universities have gained accreditation in the USA
  • Higher education is mostly funded by the State
APPROACH AND TIMETABLE FOR THE SOUTH AMERICAN SUBREGION

• Approach
  • To focus first on four countries with adequate conditions, i.e.: significant private university components; enough institutional autonomy; quality of learning concerns and/or drive; significant assessment culture
  • Concentrate initially on private leading individual universities and later, either motivate creation of ad-hoc consortia or approach existing groups through study member institutions
  • Team with education service companies and individuals with significant contacts in some countries
  • No ministry level contacts have been considered at this stage
• Timetable
  • To approach universities in March, since all four countries start the academic year in that month, as most countries in the Southern Hemisphere do. Only Colombia has a mixed calendar
  • The first semester will be taken advantage of for preparing materials and testing application readiness
  • Field application should take place in September 2017
FURTHER GENERAL REMARKS

• Some ideas:
  • Three goals of the study should be kept in mind with respect to the use of its outcomes: they allow for international comparisons; institutions within a country or group could also be contrasted; due to the design of the study, support to individual institutions, even students, will be possible on the tested competencies