For Official Use **EDU/IMHE/AHELO/GNE(2008)6**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **12-Dec-2008**

English - Or. English

DIRECTORATE FOR EDUCATION
INSTITUTIONAL MANAGEMENT IN HIGHER EDUCATION GOVERNING BOARD

EDU/IMHE/AHELO/GNE(2008)6
For Official Use

Group of National Experts on the AHELO Feasibility Study

VALIDITY AND RELIABILITY - CONSIDERATIONS FOR THE OECD ASSESSMENT OF HIGHER EDUCATION LEARNING OUTCOMES

This document was prepared by Tom Van Essen, ETS

Paris, 17-18 December 2008

The AHELO GNE is requested to:
- *DISCUSS the validity and reliability issues relevant to AHELO;*
- *TAKE NOTE of the implications of various issues on the overall validity and reliability of results, and hence on the possible uses of the AHELO feasibility study data; and*
- *In the light of these implications, ADVISE the Secretariat and AHELO experts on the way forward to address potential problems in relation to domain definition, cultural appropriateness, response types, scoring and ensuring sufficient data points.*

This document is available in PDF format only.

Contact: Karine Tremblay, Directorate for Education
[Tel: +33 1 45 24 91 82; Email: Karine.Tremblay@oecd.org]

JT03257452

English - Or. English

VALIDITY AND RELIABILITY
CONSIDERATIONS FOR THE OECD ASSESSMENT OF HIGHER EDUCATION LEARNING OUTCOMES

1. This document was prepared by Tom Van Essen. It outlines the presentation that will be made at the meeting and the questions that countries need to reflect upon during the discussion. On the basis of the discussions, a revised document will be prepared, outlining the main validity and reliability issues to take into consideration for the AHELO feasibility study, and possible strategies to address potential problems.

2. The AHELO GNE is requested to:

- DISCUSS the validity and reliability issues relevant to AHELO;
- TAKE NOTE of the implications of various issues on the overall validity and reliability of results, and hence on the possible uses of the AHELO feasibility study data; and
- In the light of these implications, ADVISE the Secretariat and AHELO experts on the way forward to address potential problems in relation to domain definition, cultural appropriateness, response types, scoring and ensuring sufficient data points.

Listening. Learning. Leading.

Validity and Reliability Considerations for the OECD's Assessment of Higher Education Learning Outcomes

OECD AHELO Meeting
Paris
December 17-18, 2008
Thomas Van Essen
Educational Testing Service

Outline

- Reliability and Validity
- Potential sources of random variance that could effect both the validity and reliability of the Assessment of Higher Education Learning Outcomes

Validity

- Most important indicator of assessment quality
- It is the extent to which
  - a test is doing the job it is supposed to do
  - inferences and actions made on the basis of test scores are appropriate and supported by evidence

Reliability And Validity

- Reliability is our ability to generalize from one assessment situation to another
  - Reliability is the consistency of the measurement
- Validity is our ability to generalize from an assessment situation to something we care about in the real world.
- An assessment can be
  - Neither reliable nor valid
  - Reliable but not valid
  - Both reliable and valid
- An assessment can **never** be
  - Valid, but not reliable

Reliability Is Necessary But Not Sufficient For Validity.

- A test may be highly reliable, but not valid for a particular purpose.
  - Very reliable measures of height would be of little use in predicting university success
- A imperfect measure of the right thing is better than a very precise measure of the wrong thing
  - An imperfectly scored essay in which candidates are asked to critique an argument is a better measure of critical thinking than 100 highly reliable multiple choice arithmetic questions

Validity and Reliability in the Context of the AHELO Pilot

- Reliability is the proportion of the total variance that is "true" (non-random) variance.
- Validity is the proportion of the true variance that is relevant to whatever it is we are trying to measure.
- What are the potential sources of random variance in the AHELO?
- And how can we reduce them?

Threats to Validity and Reliability For The AHELO Pilot

- Fuzzy domain definition
- Cultural appropriateness
- Response rates
  - sample bias
- Translation issues
- Response types
  - multiple choice
  - constructed response
- Scoring of constructed response tasks
- Insufficient number of data points to make valid and reliable inferences
- Lack of links between tasks

Domain Definition In the Subject Specific Strand

- At the third expert's meeting there was much discussion of the definition of "economics" or "biology" within national HESs
  - There is no perfect short term answer to this question
  - The Tuning approach to defining learning outcomes by discipline is most promising
  - Existing item pools or batteries should be reviewed against the Tuning framework
  - Existing item pools or batteries should be reviewed by national expert panels
    - items/tasks that are inappropriate against the Tuning framework and the national context should be dropped
      - if insufficient items remain new items created in line with Tuning framework

Domain Definition In The Generic Strand

- General agreement that “critical thinking, analytical reasoning, problem-solving, and written communication” are the key skills that should be measured.
  - But there are similar problems here as in the subject specific strands
- Are these skills the same in different linguistic and cultural contexts?
  - Is “analytic reasoning” the same in secular cultures and faith-based cultures?
  - Will “appeals to emotion” be regarded in the same way in all in all cultures?

Two Solutions: 1

- Top down
  - Develop an international framework
    - Tuning approach
    - Develop tasks that are linked to the framework
      - Disadvantages
        - » Time consuming and expensive
      - Advantages
        - » Results likely to have universal buy-in

Two Solutions: 2

- Bottom up
  - Begin with tasks
    - Review them for linguistic and cultural fit
    - Reject as many “outliers” as practically possible
      - Disadvantages
        - » very few of the tasks will please everybody or everybody equally
      - Advantages
        - » it will get done relatively quickly
        - » economical

Cultural Appropriateness

- Perhaps more of an issue in the generic strand that seeks to measure “critical thinking, analytical reasoning, problem-solving, and written communication” than in the subject specific strand
- A good critical thinking task will engage the student in an issue or problem that is relevant to his or her life situation
  - such problems are deeply enmeshed in fundamental cultural issues
- In order to make valid cross-cultural comparisons tasks should be similarly appropriate across cultures
  - A task that deals with divorce will be different in a culture where divorce is a cultural taboo than in one where it is a common and accepted everyday occurrence
- All task will need to be reviewed by a national committee for cultural appropriateness

Response Rates

- The next presentation will discuss the issue in more depth
  - but if the samples are not similar the comparisons are not valid
- We need to be sure that all samples are similarly representative
  - We must be particularly vigilant to watch out for systematic differences in samples
    - financially secure students from Institution A
    - needy students from Institution B

Translation Issues

- Detailed presentation on this issue tomorrow
  - but if the translations are not accurate the comparisons are not valid
- Issues of translation may become particularly important when we are attempting to translate complex tasks that deal with higher order skills
  - a railway schedule is a railway schedule is a railway schedule
  - a complex task in English may not be the same when it is translated into French
    - each level of cognitive and linguistic complexity represents a challenge for the translator

Response Types: Multiple Choice

- More of a threat to validity than reliability
- Advantages
  - Inexpensive to score
  - Highly reliable
  - Broad domain coverage
  - No single item has much weight
  - Linking is easier
  - Efficient in terms amount of data per unit of time testing
    - most large scale international assessments use them in combination with *short* response constructed tasks
- Disadvantages
  - Passive rather than active
  - Shallow domain coverage
    - tends to focus on lower order skills
      - recall of facts
      - procedural knowledge
  - Can drive instruction in pernicious ways

Response Types: (Complex) Constructed Response

- More of a threat to reliability than validity
- Advantages
  - Active rather than passive
  - Broad domain coverage
    - can focus on higher order skills
      - creative thinking
      - analytic reasoning
      - expressive communication
  - Sends a good signal about what we value
- Disadvantages
  - Expensive to score
  - Poor to moderate reliability
    - due to scoring issues
  - Broad domain coverage
  - Single items have great weight
  - Inefficient in terms amount of data per unit of time testing

Scoring of (Complex) Constructed Response Tasks

- Human scoring introduces variability and reduced reliability
  - especially as the tasks become more and more complex
- Tasks are scored by trained raters according to a rubric
  - our experience at ETS is that well-trained human scorers using a 0-6 scale achieve exact agreement about 60% of the time and adjacent agreement about 98% of the time
    - truncated scale will lead to higher agreement rates
    - extended scales will lead to lower agreement rates
  - Higher levels of exact agreement are not necessarily a good thing
    - raters may be evaluating on length or some other construct irrelevant variable rather than quality of argument
- (Talk to me about machine scoring during the cocktail hour)

Checks on Scoring In The Transnational Context

- Raters from different cultural and linguistic groups may differ in systematic ways
- For the feasibility study complex tasks should be "over scored"
  - each task should be scored by at least two raters
- Scoring should be designed and analyzed so as to allow comparisons between
  - Native speakers of language A vs. native speakers of language A
  - Native speakers of language A vs. bilingual speakers of languages A and B
  - Native speakers of language B vs. native speakers of language B
  - Native speakers of language B vs. bilingual speakers of languages B and C
  - Native speakers of language C vs. native speakers of language C
  - Native speakers of language C vs. bilingual speakers of languages C and A
- All groups should use the entire score range
- All groups should achieve similar level of exact and adjacent scoring

Crosstabs Across Groups Should Be Similar

1st Score

2nd Score / Score Category	0	1	2	3	4	5	Total N
0	11	7	2	0	0	0	*20*
1	5	29	10	5	0	2	*51*
2	2	14	53	21	0	9	*99*
3	1	1	19	68	31	11	*131*
4	5	3	16	20	112	36	*192*
5	9	0	5	3	14	29	*60*
Total N	*33*	*54*	*105*	*117*	*157*	*87*	*553*

Sufficient Number of Data Points

- Conflicting imperatives
  - testing time
  - construct representation
  - reliability
  - cost

- Since the AHELO is interested in performance at the department or school level, fewer data points per individual are needed than for an assessment that would issuing individual scores fewer data points
- What is the appropriate trade off ?

Insufficient Linkages Between Tasks

- If all we know is that student A gets a 5 on the “blue” task and student B gets the a 6 “red” task can we say that student A is better?
  - We don’t know what they measure
  - We don’t know the scale
  - We don’t now the measurement error
  - We don’t now anything about where they come from
- The feasibility study should be designed so that all students get common items or tasks that are linked to each other on some way
- This will allow for meaningful comparisons
- This will allow for calculations of reliability and validity

Questions?
Discussion?