

**DIRECTORATE FOR EDUCATION
INSTITUTIONAL MANAGEMENT IN HIGHER EDUCATION GOVERNING BOARD**

Group of National Experts on the AHELO Feasibility Study

ANALYSIS AND REPORTING DESIGN

9th meeting of the AHELO GNE

Paris, 19-20 March 2012

This document was prepared by the ACER Consortium.

The AHELO GNE is expected to COMMENT and DISCUSS the analysis and reporting design.

Contact:
Consortium: ahelo@acer.edu.au
OECD Directorate for Education: Diane.Lalancette@oecd.org

JT03317570

Complete document available on OLIS in its original format

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

TABLE OF CONTENTS

INTRODUCTION	3
DATA AND FILE PREPARATION	3
Data cleaning and verification.....	3
Data file building.....	4
SAMPLING ANALYSIS AND WEIGHTING	5
SCALING	6
VALIDITY ANALYSES.....	7
CONTEXTUAL ANALYSES.....	9
DATA PRODUCTS AND WRITTEN REPORTS	9
Database and Codebooks.....	10
Compendia	10
Technical Report	10
Institution Reports.....	10
Study Report.....	10

INTRODUCTION

1. The Assessment of Higher Education Learning Outcomes (AHELO) Feasibility Study is a major OECD project. Its objective is to determine a robust approach to measuring learning outcomes in ways that are valid across cultures and languages, and across the diversity of institutional settings and missions.
2. This document provides an overview of designs for analysis and reporting for AHELO. It covers data file preparation, sampling analysis and weighting, scaling, validity analyses, contextual analyses, and the production of data products and written reports.
3. This analysis and reporting design is distinct to the AHELO Analysis Plan—the latter being the study’s evaluation architecture. A further related document is the AHELO Data Access, Use and Reporting Policy.
4. In principle, though not necessarily within the scope of the AHELO Feasibility Study, analysis and reporting has to satisfy three audiences: policymakers and institutions who want reasonable detailed statistical reports; the public and broad stakeholders who want results which are easy to read and digest; and researchers who will demand high levels of detail, rigor and proof.
5. AHELO is designed to produce international reports and also reports for participating higher education institutions (HEIs). Many analyses are conducted at these two levels. Several validity analyses are also pitched at the national or system level, however, given that work is stratified in this way for translation and adaptation. No national results or reports will be produced.

DATA AND FILE PREPARATION

Data cleaning and verification

6. Data will be prepared and files produced to adhere to international specifications, enable linkage across different instruments, and ensure accurate and consistent storage of information. Data cleaning will include:
 - a) file structure and valid range checks;
 - b) identification-variable (ID) cleaning;
 - c) between-file linkage checks;
 - d) implementing flow and filter edits;

- e) cleaning of background inconsistencies;
- f) re-arranging the file structure towards data analysis and application of general cleaning rules; and
- g) quality control cleaning.

7. Deviations and problems will be labelled with a unique problem number and a description of the problem. Appropriate action will be taken by the AHELO Consortium. If problems are identified that cannot be automatically adjusted they and proposed solutions will be reported to the responsible NPM.

8. Verification will be conducted to ensure that elements and values in the files link back to empirical responses. Reference will be made to original source files harvested by online collection software. This validation may require consultation with NPMs and Lead Scorers.

Data file building

9. As part of instrument and system development the AHELO Consortium has prepared detailed data file structures and element specifications to prepare for data which will be received from systems, institutions, faculty and students in 2012. The data file structures identify all elements of student, faculty, institution and national data which will be recorded and specifies the format in which data will be recorded and its precise characteristics.

10. By way of summary, in addition to various lookup and linking tables four flat unit-record data files will be prepared reflecting the multilevel study design:

- a) Country Data File (CDF)—about 20 elements mainly sourced from existing policy documentation;
- b) Institution Data File (IDF)—about 90 elements including management elements and data sourced using the Institution Context Instrument (ICI);
- c) Faculty Data File (FDF)—about 50 elements, including management elements, sampling elements, and data sourced using the Faculty Context Instrument (FCI); and
- d) Student Data File (SDF)—about 250 elements total, including management elements, sampling elements, and data sourced using test instruments and the Student Context Instrument (SCI), which a breakdown of elements as follows:
 - i. management elements (about 20 elements, 8% of SDF);
 - ii. sampling elements (about 15 elements, 6% of SDF);
 - iii. Generic Skills constructed response tasks (3 elements, 1% of SDF);
 - iv. Generic Skills MCQs (about 40 elements, 16% of SDF);
 - v. Economics constructed response tasks (about 15 elements, 6% of SDF);
 - vi. Economics MCQs (about 50 elements, 21% of SDF);
 - vii. Engineering constructed response tasks (about 20 elements, 8% of SDF);

- viii. Engineering MCQs (about 40 elements, 16% of SDF); and
 - ix. Student Context Instrument (about 40 elements, 16% of SDF);
11. The use of online collection mechanisms helps ensure the collection of consistent and valid data. Planned and unplanned missing data still exists, however, as do additional elements for particular systems. The following deviations from the international file structure will be identified for each system and HEI:
- a) international variables omitted;
 - b) national variables added; and
 - c) notable manifestations of missing data.
12. Missing data will be treated using a variety of marginal and conditional single imputation procedures. These will be documented.
13. Data files will be validated using standard procedures. This includes univariate and bivariate descriptive statistical analysis, the imputation of flags and non-response codes, and the addition of various dummy-coded management variables. Verification work will be conducted to ensure that the data in defined files matches that collected from NPMs or online instruments.
14. Once raw data files have been prepared a series of derivative files will be produced to support various forms of analysis and reporting. These files will contain different amounts of aggregation and composite variables. All files will be archived in SPSS and straight text formats.
15. Descriptive summaries of all variables will be produced for each system and institution. National Centres and NPMs will be asked to review these summaries and note any objections to data or files.

SAMPLING ANALYSIS AND WEIGHTING

16. In parallel with the preparation of the data file a series of marker variable analyses will be conducted to check coverage and representativeness of secured data. This analysis will take stock of response rates by institution and strand. Results of these analyses will be checked with NPMs where they deviate from normal.
17. The sampling and analysis teams will collaborate on the computation of student-based weights for each strand and institution. These weights will be adjusted for student non-response within institution or program (as required by the sampling plan adopted in each institution and strand).
18. Even though the approach to variance estimation does in principle depend on the sampling approach adopted within HEIs, to ensure a consistent approach to analysis across HEIs and across systems it would likely be appropriate to develop replication weights and apply a replication method to variance estimation for all HEIs and strands. For most simple statistics (such as means, totals and ratios of totals) the value of the sampling error estimated by replication will be identical to the value obtained using non-recursive formulae.

19. A review of sampling outcomes will be conducted that summarises response characteristics, sampling variance, and clustering and homogeneity. This review will also explore in detail the processes and outcomes of various adjudication procedures.

SCALING

20. The Consortium will use scaling methodology based on item response modelling. This methodology is widely used in international studies in the field of education and enables researchers to assess scaling characteristics. The analysis will place special emphasis on the review of the cross-cultural comparability of all instruments used in this study.

21. More specifically, the Consortium will use the Rasch model in its general form as implemented in the ACER ConQuest software to analyse the AHELO Feasibility Study assessment data. This analytic method will be used because:

- a) of all available item response theory models, it provides the strictest assessment of psychometric validity;
- b) it supports the construction and validation of meaningfully described proficiency scales, which are taken as a requirement for the useful reporting of the AHELO Feasibility Study assessment data;
- c) it has been widely generalised to deal with the range of analytic requirements of complex cross-national studies similar to the AHELO Feasibility Study (such as exploring and controlling for scorer effects and item position effects, and supporting multidimensional scaling); and
- d) it also supports equating tests for the purposes of maintaining and monitoring item sampling and trends.

22. Scaling will proceed in three stages:

- a) national calibrations, in with the scaling will be replicated across national contexts to ascertain the stability of item parameter estimates;
- b) international calibration, which will be based on full international data or replicated across several subsamples drawn from the international data; and
- c) individual ability estimate generation, in which plausible values are estimated for each student completing a test.

23. The performance of items will be analysed and reviewed during scaling. The effect of item orderings and rotations will be noted and reviewed. Any item deletions will be recorded.

24. Assessment results will be scaled onto a standard metric. This metric will have an international mean of 500 and international standard deviation of 100. As possible given time and performance characteristics, the measured variable will be analysed to identify thresholds that distinguish different

levels of performance. If possible, these will be content analysed to define progressions of increasing capability.

25. Student ability estimates will be computed using multiple imputation techniques, and for each student five plausible values will be obtained using ACER ConQuest. For generating plausible values all available background variables will be used as conditioning variables. This will increase the precision of estimates and allow the analysis of relationships between ability estimates and background variables. Using plausible value methodology to obtain individual estimates is a standard approach in the national and international assessment studies. Simulation studies have clearly demonstrated the superior estimates obtained from plausible values, in particular when analysing them in conjunction with school or student background variables.

26. The results for the five plausible values can be used to obtain an estimation of measurement error. Estimates of sampling and measurement error will be combined to obtain final standard errors for all performance statistics reported for the sample.

VALIDITY ANALYSES

27. As a large and innovative study it is necessary to analyse the validity of several facets of AHELO method and data. These analyses focus on the validity of instruments and applications of use, and on whether results generalise across reporting and contexts.

28. In addition to item response modelling a range of other classical analyses will be conducted to generate reliability and validity statistics, and test the efficiency of alternate scoring methods. Item-total statistics will be generated for each national group. A series of reliability generalisability (RG) studies will be conducted to review whether the errors of measurement are stable across contexts. Reliability estimates will be produced for the student and institutional levels.

29. Item fit to the measurement dimension will be assessed using a range of item statistics. The weighted mean-square statistic (infit), a residual based fit statistic, will be used as a global indicator of item fit. Weighted infit statistics will be reviewed both for item and step parameters. The analysis of item fit and the estimation of item parameters will be carried out with the ACER Conquest.

30. Item response modelling will be used to assess the ‘targeting’ of the test to respondent cohorts. This involves checking whether the distribution of item difficulty maps well against the distribution of respondent capability.

31. In addition to this, item characteristic curves (ICC) will be generated for every item, which provide a graphical representation of item fit across the range of student abilities for each item (including dichotomous and partial credit items). The functioning of the partial-credit scoring guides will further be reviewed through investigation of the proportion of responses allocated to each response category and the differences in mean abilities of students by response category.

32. In international studies the issue of cross-contextual validity of instruments is of crucial importance in order to ensure that data are comparable across institutions and languages. In the AHELO Feasibility Study, such analyses will play a central role, for they will help identify the extent to which the

assessments have been successfully generalised cross-nationally, cross-culturally, cross-linguistically and cross-institutionally.

33. The cross-contextual validity of the test items will also be explored by assessing differential item functioning (DIF) (for groups that have sufficient sample). Specifically, IRT will be used to detect variance of item parameters across contexts. Such variance indicates that groups of students with the same ability have different probabilities of responding correctly to an item. This is commonly referred to as ‘item bias’ as it indicates that the probability of successful performance is a function of group membership as well as individual ability.

34. As an adjunct to performance data, audit data captured through the online delivery mechanism will be analysed in detail to determine (as possible) whether the technology mediated the performance of different student groups, and factors such as the time students spend on items and the order in which students choose to interact within items.

35. AHELO instruments use a range of item types. Psychometric analyses will check the synergies between these different item types. Analyses will be conducted to explore whether there are any off-dimensional interactions between assessment items and instruments and student or institutional groups. These analyses will take account of item content and difficulty as well as respondent characteristics.

36. Unplanned item-level non-response will be analysed in detail to identify salient response patterns and measurement disturbances. The extent to which missing student responses are due to problems with test length (‘not reached items’) will be reviewed to assess the appropriateness of test length. Unreached responses will be defined as all consecutive missing values starting from the end of the test except the first missing value of the missing series, which will be coded as ‘missing’. An analysis of skipped non-response will also be conducted to spotlight problematic items.

37. Analyses will be conducted of the generalisability of constructed response task data across national and linguistic contexts. These analyses will review:

- a) whether the statistical distribution of scores from subjectively scored constructed response tasks varies (in terms of effect size units) across national, institutional and disciplinary contexts;
- b) the extent to which construct generalisability across institutions, and particularly national/linguistic contexts, implies that scoring rubrics have been interpreted invariantly;
- c) inter-rater reliability, using variance decomposition analyses based on cross-rater reliability statistics collected via online systems during fieldwork, and reviewing whether standards were met; and
- d) consistency of scoring and scoring outcomes across national contexts, established by cross-scoring translated response tasks or capturing agreement between Lead Scorers.

CONTEXTUAL ANALYSES

38. After the Consortium has confidence in the quality of the assessment data and scores statistical analyses will be conducted for exploratory and explanatory purposes. Results will be reported in international, institutional and technical data products and written reports.

39. Review of percentages across categories will be computed and will provide information about the possible skew of items and the amount of missing responses. Analytical preference will be given to items that have sufficient percentages in each category. Distributions of core background variables will also be compared with those from previous surveys both the international and national level.

40. The context data will be used to perform general descriptive analysis in which cognitive outcomes are disaggregated by major student groupings and institutional characteristics. These analyses will be replicated across strands and within institutions.

41. The Consortium will employ multilevel modelling to glean a more nuanced understanding of the data and test for differential performance across contexts. This approach will help estimate how performance is related to key factors at various levels of nesting, students, departments, institutions and systems. Such modelling would most directly address the question of the nature of contextual effects. Contextual and confounding variables are accounted for at each level of nesting while the design effect of clustering is taken into account for the estimation of standard errors. These models not only describe the relationships between context and outcomes, but they also show how some effects vary from institution to institution, and ultimately enable the team to identify sources of variation. By comparing relationships between outcomes and background contexts they facilitate a kind of concurrent validation.

42. For contextualisation and reporting, as opposed to criterion validation, comparison will be made to available and relevant benchmark scores for particular strands and respondent subgroups. Based on feedback from AHELO NPMs and GNE benchmark group scores will be computed for:

- a) ALL HEIs internationally in each strand; and
- b) SELECTED HEIs internationally in each strand, with institutions selected using contextual matching criteria.

DATA PRODUCTS AND WRITTEN REPORTS

43. Several resources will be produced to support data analysis and enable effective international dissemination of the results from the AHELO Feasibility Study. In consultation with AHELO's Expert Groups, the Consortium will produce:

- a) Database and Codebooks;
- b) Compendia;

- c) Technical Report;
- d) Institution Reports; and
- e) Study Report.

Database and Codebooks

44. The database will include all student scores, all final weights and replicate weights for sampling variance computation any context composite indices derived from the questionnaires, together with students' responses to the questionnaire and the test questions. This database will allow the OECD Secretariat, the AHELO GNE, NPMs and institutions to conduct their own further analyses. Data will be provided in ASCII format and in STATA, SAS and SPSS format. In addition, codebooks will be provided for all databases that list the variable names, variable labels, format, column position in the ASCII data files, categories and labels.

Compendia

45. The Compendia will be prepared and made available in Adobe PDF, MS Word and MS Excel formats. Compendia will be developed that include a set of tables showing statistics for every item in the questionnaires, and the relationship of background variables with performance. The tables will show the percentage of students per category of response and the average performance by assessment and domain for the groups of students in each category.

Technical Report

46. The Feasibility Study Technical Report will summarise all technical aspects and standards of the AHELO Feasibility Study. It will clearly describe all data and statistical conventions and approaches applied in the study, information on matters of test and questionnaire design, field operations, sampling, data adjudication and quality control mechanisms, methodologies used to analyse the data and other technical features of the project will be described at a level of detail that allows researchers to understand and replicate its analyses.

Institution Reports

47. These reports will provide HEIs with information on their students. The reports will consist of the full dataset for a given institution as well as the institutional performance profile including benchmarks of other participating higher education institutions in a way that does not allow benchmark institutions to be identified. A code of practice will be developed to provide guidelines on how AHELO Feasibility Study data should be used, as well as provide ideas on ways to use the data and results most effectively.

Study Report

48. The final report will include the methodological and technical questions raised by an international AHELO – including issues of domain definition, conceptual assessment frameworks, validity of instruments, translation and cultural adaptation, field implementation, scoring, scaling and reliability of results, data analysis; issues that arose during implementation and in the analysis of results; and conclusions on the scientific and practical outcomes of the feasibility study as well as guidance for the longer-term development of an AHELO should the initiative be taken forward.