Chapter 17: INTERNATIONAL DATA PRODUCTS

After the data processing and data analysis, a series of data products were delivered to the OECD. These included public use data files and codebooks, compendia tables, and the PISA Data Explorer, a data analysis tool. These data products are available on the OECD website *(http://www.oecd.org/pisa/pisa-for-development/*). The International Association for the Evaluation of Educational Achievement (IEA) IDB Analyzer was configured to work with PISA-D data and can be downloaded at (*https://www.iea.nl/data*).

PUBLIC USE FILES

The international public use data files combine all international reportable countries into one file and includes an approved set of international variables that are common to all countries. Each national database includes approximately 1,000 common variables for student cognitive and context questionnaire assessments and approximately 400 school and teacher variables. A subset of these were included in the public use data files, made available on the OECD website at *(http://www.oecd.org/pisa/pisa-for-development/*).

Variables excluded or suppressed for some or all countries

The public use data files include only a subset of the information available in the master databases for each country. The public use data files do not include any data collected using national adaptations and extensions. They include only data that were collected or derived across all countries. Further, a sizable number of variables were excluded in consultation with the OECD Secretariat because they i) have little or no analytical utility, ii) were intended for internal or interim purposes only, iii) relate to secure item material, or iv) include personally identifiable data, or at least data that may increase the risk of unintended or indirect disclosure.

The groups of variables excluded from the public use data files are:

1. direct, indirect, and operational identifiers for respondents
2. certain context questionnaire variables, especially detailed free-text entry items
3. all national adaptations and extensions in the context questionnaire
4. original scale score values (theta) before standardisation to an international metric

As discussed in Chapter 10, countries were given the option of suppressing variables in the public use files. Suppression of variables was approved when data presented a risk to student, school, and/or teacher anonymity, or for technical errors that could not be resolved by data contractors. Suppressed data are represented in the database by means of missing codes.

File names and content

There are four public use data files: the student questionnaire data file (which also includes estimates of student performance data), the school questionnaire data file, the teacher questionnaire data file, and the cognitive item data file.

Data files are provided in both SAS and SPSS formats. The files include:

- **Student questionnaire data file (PUF_COMBINED_CMB_STU_QQQ.zip):** This file includes ID variables, the Student Context Questionnaire responses, student scale and derived variables, plausible values (Reading, Math, and Science), and overall and replicate weights.
- **School questionnaire data file (PUF_COMBINED_CMB_SCH_QQQ.zip)**: The school questionnaire data file includes ID variables, school questionnaire data, school questionnaire scale and derived variables, and an overall school weight.
- **Teacher questionnaire data file (PUF_COMBINED_CMB_TCH_QQQ.zip)**: The teacher questionnaire data file includes ID variables, teacher questionnaire data, and teacher questionnaire scale and derived variables.
- **Student cognitive item data file (PUF_COMBINED_CMB_STU_COG.zip):** The cognitive data file includes ID variables and raw, scored, and coded items.

Variables used in sampling, weighting, and merging

The variable *STRATUM* is included to differentiate sampling strata. The variable is created as a concatenation of a three-letter country code, a two-digit region identifier, and a two-digit original stratum identifier.

The variable *W_FSTUWT* contains the final student population weight, and the variables *W_FSTURWT1* through *W_FSTURWT80* contain the replicate weights used to determine sampling error. The variable *SENWT* is a normalised (senate) weight variable based on *W_FSTUWT* for analyses of student performance across a group of countries where contributions from each of the countries in the analysis are desired to be equal regardless of their population or sample size. The senate weight makes the sum of the weights of each country be 5,000 to ensure an equal contribution by every country in the analysis when countries are analysed combined. This weight is only applicable to the student variables that do not contain missing values. Its application to other variables might be compromised by its dependence on the patterns of missing data.

The student and teacher data files can each be merged to the school data file using the variable *CNTSCHID*. *CNTSCHID* is the combination of the three-digit country code and a randomised five-digit number, making it unique across all countries. When merging the student questionnaire data file with the student cognitive item data file, use CNTSTUID. When merging student level data with school level data, use CNTSCHID. When merging teacher data with school level data, use CNTSCHID. CNTSCHID, CNTSTUID (in the student file), and CNTTCHID (in the teacher file) have had their values randomised from the original order received during country submission while

still retaining the original student-to-school and teacher-to-school connection.

Missing code conventions

The data may include up to six MISSING categories:

1. Missing/blank (“.” in SAS; blank or “SYSMIS” in SPSS)—In the cognitive data, it is used to indicate the respondent was not presented the question according to the survey design or ended the assessment early and did not see the question. In the questionnaire data, it is only used to indicate that the respondent ended the assessment early or, despite the opportunity, did not take the questionnaire.

2. No response/omit (“.M” in SAS; “9/99/999/…” in SPSS)—The respondent had an opportunity to answer the question but did not respond.

3. Invalid (“.I” in SAS; “8/98/998/…” in SPSS)—Used to indicate a questionnaire item was not conforming to the expected response (e.g., the respondent indicated more than one choice for an exclusive-choice question).

4. Not applicable (“.N” in SAS; “7/97/997/…” in SPSS)—The response could not be determined due to a printing problem or torn booklet. It can also be used to indicate that an item for a specific country was deleted during item calibration.

5. Not reached (“.R” in SAS; “6/96/996/…” in SPSS)—Used in the cognitive data to indicate that a student may have run out of time and did not reach the item in question. It is done by recoding each student’s successive “No response/Omit” values at the end of each cluster. This code is assigned during processing.

6. Valid skip (“.V” in SAS; “5/95/995/…” in SPSS)—The question was not answered because a response to an earlier question directed the respondent to skip the question. This code is assigned during data processing.

Codebooks for the PISA-D public use data files

Included with the PISA-D Main Survey data products is a set of data codebooks in Excel format. The data codebook is a printable report containing descriptive information for each variable contained in a corresponding data file. The codebooks report frequencies and percentages for all variables that employ a value scheme for cognitive and questionnaire variables, as well as those that have been derived and/or added during data cleaning. The codebooks are available on the OECD website *(http://www.oecd.org/pisa/pisa-for-development/*).

The information is displayed with variable names, variable labels, values, and value labels. Other metadata is provided, such as variable type (e.g., string or numeric) as well as precision/format. Additionally, the codebooks contain a range of values (minimum and maximum) for those numeric variables that do not employ a value scheme.

Codebooks for the main files are contained in four separate worksheets (**Codebook_CMB.xslx**):

1. Student—Student questionnaire data include parent, educational career, and information communication and technology questionnaire data)
2. School—School questionnaire data
3. Cognitive—Student cognitive data for Reading, Mathematics, and Science
4. Teacher—Teacher questionnaire data

DATA ANALYSIS AND SOFTWARE TOOLS

Standard analytical packages for the social sciences and educational research do not readily recognise or handle the complex PISA sample and assessment design. This gap is filled by the two software tools made available to assist database users to access and analyse PISA data and produce basic outputs: the PISA-D Data Explorer (PDX) and a microdata analyser called the IEA IDB Analyzer. Each of these tools addresses a slightly different set of needs. While the PDX is a web-based application that allows relatively easy and publication-ready access to basic estimates of means, totals, and proportions, the IEA's IDB Analyzer, used in conjunction with the public use files (PUFs), allows unit record access to the public use database and the opportunity to conduct analysis offline, derive additional variables, and produce various estimates for further use and reporting. The PDX and IEA's IDB Analyzer are described in turn in the remainder of this chapter.

PISA-D Data Explorer (PDX)

The PDX is a web-based application that allows the user to query the OECD-hosted, secure, PISA-D International Database via a web browser. In addition to PISA-D microdata, the PDX database contains the PISA 2015 microdata for variables that are common to both surveys, PISA-D, and PISA 2015. The PDX is available on the OECD website *(http://www.oecd.org/pisa/pisa-for-development/*) and provides access to a secure PISA-D database (protected by the OECD firewalls and security mechanisms) to navigate, analyse, and produce report quality tables and graphics.

Because certain variables that are included in the PUF for secondary analysis are not informative or usable as part of the PDX, they are not included in the PDX database. The variables not included in the PDX relate to the individual cognitive items and response process information.

In the PISA-D Data Explorer, the OECD average is created from data for the 35 countries that were members of the OECD at the time of release. The PISA-D average is created from data for all of the countries in the PISA-D study.

The PDX can be used to compute a diverse range of statistics including, but not limited to, means, standard deviations, standard errors, percentages by subgroup, percentages by performance levels, and percentiles. All statistics are computed taking into account the sampling and assessment design. In addition, the PDX has the capability of conducting significance testing between statistics from different groups and displaying the results in graphical form. Results from the PDX can be directly exported and saved to Microsoft Word, Microsoft Excel, PDF, and HTML formats.

Because it is web-based, and processing takes place on a central server, the PDX can be accessed and used with computers that meet fairly simple requirements. The user's computer is used only to create a request or data query, deliver the request to a central server where processing takes place, and then receive and display back the results in a user-friendly format.

A typical query consists of the user selecting the domain(s), jurisdiction(s), and variable(s) of interest. Then the user proceeds to select the statistics of interest and format the table. Statistics are calculated for each of the subgroups defined by the variable or variables, for one variable at a time or in cross-tabulation mode. In addition, the user is able to collapse categories for each of these variables and use the collapsed categories in the analysis. All statistics are calculated using weighted data, with their corresponding standard errors taking into account sampling and measurement uncertainty. The user has the option to select whether the standard errors are displayed in the table or not, as well as the precision with which the statistics are displayed. The results can then be displayed in a table or in a graphic.

Regardless of whether the results are displayed in a table or graphic mode, the results can be saved or exported for further post-processing or for inclusion in an external document.

A significance test module allows the user to specify significance testing to be done between subgroup means, percentages, and percentiles, while implementing necessary adjustments that take into account the sample and test design. Significance test results can be displayed in table or in graphic format.

Table results can be easily exported to be manipulated using spreadsheet software, allowing the user to customise the titles and legends of the tables, and to do any required post-processing. Likewise, the graphic results can also be exported to be included in documents for future display and use in reports and presentations.

The web application is compatible with many widely used browsers including Internet Explorer 7 and higher, Firefox 3.0 and higher, Google Chrome, and Safari. Target screen resolution is 1024 x 768. Users should enable JavaScript and pop-ups in their browsers and install Adobe Flash Player 9.0.115 or higher.

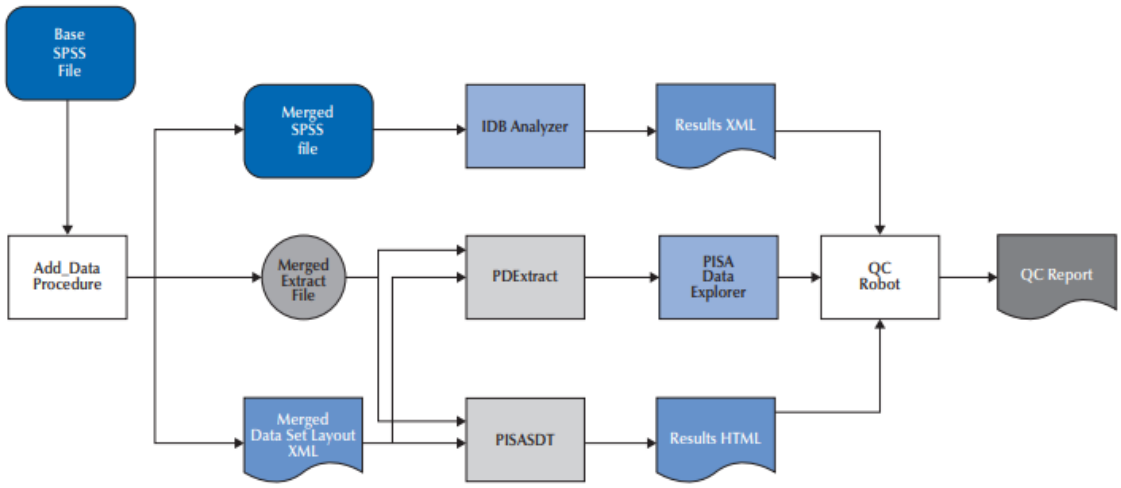
Population and quality check of the PISA Data Explorer

The process to populate the PISA Data Explorer database and confirm the results it produces is summarised in Figure 17.1. This process was applied separately to the data from each country.

The Base SPSS File contained the data as forwarded to the appropriate country for its analysis and reporting.

■ Figure 17.1 ■

PISA database population and quality control



The Add_Data procedure performed two functions. The first was conditional on whether a country provided supplemental data that was collected or derived and merged these data with the Base file. The second function created two files from the enhanced Base file: an ASCII text rectangular file containing the data values extracted from the Base file and an XML file containing information about the extracted data variables (location, format, labels). This Data Set Layout (DSL) XML is structured in a proprietary ETS schema.

The PDExtract program used the information from an input parameter file to process the data from the Extract file and metadata from the DSL file to produce a series of text files suitable for loading into the appropriate tables in the PISA Data Explorer (PDX) database. The program also produced a SQL script that is customised for performing the loading of these tables and contains a procedure for forming the data tables used by the PDX.

The PISASDT program also used the information from an input parameter file as well as a list of data variable names to calculate and produce summary data tables (SDT)—one analysis for each scale score. Each table in the analysis was a one-way tabulation of various statistics for each category of a given variable. The statistics pertained to a scale score and include percentage, average score, and percentages within the benchmark levels. Each statistic was accompanied by the standard error estimate, degrees of freedom, number of cases on which the statistic was based, and number of strata on which the standard error was based. All of these results were stored in an HTML document in full precision. This document may be viewed with any of the popular internet browsers when accompanied by the appropriate Cascading Style Sheet (CSS) document, which ETS provided. The document may also be parsed or translated to produce Excel workbooks and report quality tables, among others.

In the QC Robot procedure, the Results HTML document from the PISASDT program was used to generate analysis requests for the PDX, one for each variable, and the results returned from the PDX were compared with those in the HTML document. The results of these comparisons were

posted to the QC Report document, where differences above specified criteria were flagged and subsequently examined.

The only statistics that can be reported in the PDX which cannot be calculated by the PISASDT program are the percentiles. Because the calculation of the percentiles within the PDX uses more resources than the other statistics, only a subset of critical variables was selected for quality-assurance analysis. The Analyzer reads data from the Base SPSS file, uses SPSS macros to calculate the desired percentile statistics, and writes the results to an XML file. The QC Robot procedure processed this XML file in the same way as the HTML file from the PISASDT program and added the comparison results to the QC Report file.

Prior to the first execution of the procedure described above, the Analyzer and the PISASDT programs were extensively calibrated with each other to ensure that the Merged SPSS and Merged Extract files were isomorphic and produced identical results for the statistics common to both programs.

IEA's International Database Analyzer

The IEA's International Database Analyzer (IDB Analyzer) is an application developed by the IEA-Hamburg that can be used to analyse data from most major large-scale assessment surveys, including those conducted by the OECD, such as PISA and PISA-D. Originally designed for international large-scale assessments, it is also capable of working with national assessments such as the U.S. National Assessment of Educational Progress (NAEP).

The IDB Analyzer creates SPSS or SAS syntax that can be used to perform analysis with these international databases. The syntax takes into account information from the sampling design in the computation of sampling variance, and handles the plausible values. The code generated by the IDB Analyzer enables the user to compute descriptive statistics and conduct statistical hypothesis testing among groups in the population without having to write any programming code.

The IDB Analyzer is licensed free of cost, not sold, and is for use only in accordance with the terms of the licensing agreement. A complete copy of the licensing agreement is included in the Appendix of the Help Manual of the IDB Analyzer.

The analysis module of the IDB Analyzer provides procedures for the computation of means, percentages, standard deviations, correlations, and logistic and linear regression coefficients for any variable of interest overall for a country, and for specific subgroups within a country. It also computes percentages of people in the population that are within, at, or above benchmarks of performance or within user-defined cut points in the proficiency distribution, percentiles based on the achievement scale, or any other continuous variable.

The analysis module can be used to analyse data files from PISA-D. The following analyses can be performed with the analysis module:

1. Percentages and means: Computes percentages, means, design effects and standard deviations for selected variables by subgroups defined by the user. The percent of missing responses is included in the output. It also computes t-test statistics of group mean differences taking into account sample dependency.

2. Percentages only: Computes percentages by subgroups defined by the user.

3. Linear regression: Computes linear regression coefficients for selected variables predicting a dependent variable by subgroups defined by the user. The IDB Analyzer has the capability of including plausible values as dependent or independent variables in the linear regression equation. It also has the capability of contrast coding categorical variables (dummy or effect) and including them in the linear regression equation.

4. Logistic regression: Computes logistic regression coefficients for selected variables predicting a dependent dichotomous variable, by subgroups defined by the user. The IDB Analyzer has the capability of including plausible values as independent variables in the logistic regression equation. It also has the capability of contrast coding categorical variables and including them in the logistic regression equation. When used with SAS, the user can also specify multinomial logistic regression models.

5. Benchmarks: Computes percent of the population meeting a set of user-specified performance or achievement benchmarks by subgroups defined by the user. It computes these percentages in two modes: cumulative (percent of the population at or above given points in the distribution) or discrete (percent of the population within given points of the distribution). It can also compute the mean of an analysis variable for those at a particular achievement level when the discrete option is selected. New in 2016 is the computation of group mean and percent differences between groups taking into account sample dependency.

6. Correlations: Computes correlation for selected variables by subgroups defined by the grouping variable(s). The IDB Analyzer is capable of computing the correlation between sets of plausible values.

7. Percentiles: Computes the score points that separate a given proportion of the distribution of scores by subgroups defined by the grouping variable(s).