Chapter 19
International Data Products

Following the data processing and data analysis, 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/*). The IEA's IDB Analyzer was configured to work with PISA data and can be downloaded from http://www.iea.nl.

PUBLIC USE FILES

The international public use data files combine all international reportable countries into one file and include an approved set of international variables that are common to all countries. The public use data files contain approximately 4700 variables for student cognitive and background questionnaire assessments, approximately 3100 variables for financial literacy student cognitive and background questionnaire assessments, and approximately 546 variables for school and teacher background questionnaire assessments. The public use data files and corresponding documentation are available on the OECD website at *http://www.oecd.org/pisa/*.

Variables excluded or suppressed for some or all countries

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

- direct, indirect, and operational identifiers for respondents
- all national adaptations and extensions in the BQ
- 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. Suppressed data are represented in the database by means of missing codes.

File names and content

There are eight public use data files. Data are provided in both SAS and SPSS formats within each file. The file names and content are described in Table 19.1.

Table 19.1 : PISA 2018 Public Use Data Files

FileName	File Description	StQ	StQ (FLit)	ECQ	PQ	ScQ	TQ	ICTQ	PVs (RSM)	PVs (RM)	PVs (FLit)	IR	TP	Wgts	RWgts
CY07_MSU_STU_QQQ	Student Questionnaire (Main Sample)	Yes		Yes	Yes			Yes	Yes					Yes	Yes
CY07_MSU_STU_COG	Student Cognitive (Main Sample)											Yes	Yes	Yes	Yes
CY07_MSU_STU_TIM	Student Questionnaire Timing (Main Sample)												Yes		
CY07_MSU_FLT_QQQ	Student Questionnaire (Financial Literacy)	Yes	Yes							Yes	Yes			Yes	Yes
CY07_MSU_FLT_COG	Student Cognitive (Financial Literacy)											Yes	Yes	Yes	Yes
CY07_MSU_FLT_TIM	Student Questionnaire Timing (Financial Literacy)												Yes		
CY07_MSU_SCH	School Questionnaire					Yes								Yes	
CY07_MSU_TCH	Teacher Questionnaire						Yes								

The description of the contents of the datafiles is as follows:

Key	Description
StQ	Student background questionnaire responses and derived variables
StQ (FLit)	Student Financial Literacy background questionnaire responses and derived variables
ECQ	Early Career Questionnaire responses and derived variables
PQ	Parents Background Questionnaire and derived variables
ScQ	School Background Questionnaire and derived variables
TQ	Teacher Background Questionnaire and derived variables
ICTQ	Information and Communication Technology Questionnaire and derived variables
PVs (RSM)	Plausible Values for Reading, Math and Science (Main Sample)

PVs (Read)	Plausible Values for Reading Subscales (Main Sample)
PVs (RM)	Plausible Values for Reading and Math (Financial Literacy Sample)
PVs (FLit)	Plausible Values for Financial Literacy
IR	Cognitive Item Reponses (raw and scored responses)
TP	Timing and process data (number of actions, total time, time to first action, etc.)
Wgts	Overall sampling weight
RWgts	Replicate sampling weights

Variables used in sampling, weighting and merging

The following are sampling and weighting related variables included in the data files

STRATUM : 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.

SENWT: This is a normalised (senate) weight variable 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 is a transformation of the full sampling weight adding to 5 000 within each country. This weight is only useful for student variables that do not contain missing values. Its application to other variables might be affected by its dependence on the patterns of missing data.

W_FSTUWT: This is the full student sampling weight variable to be used when analyzing student level data.

W_FSTURWT1 to W_FSTURWT80: These are the replicate weights to be used in the calculation of sampling variance.

WVARSTRR: RANDOMIZED FINAL VARIANCE STRATUM (1-80)

UNIT: RANDOMLY ASSIGNED UNIT NUMBER

W_SCHGRNRABWT: Grade nonresponse adjusted school base weight

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. *CNTSCHID*, *CNTSTUID* (in the student file), and *CNTTCHID* (in the teacher file) have had their values randomised from the original order received from each country/economy, while still retaining the original student to school and teacher to school connection.

Student level data files can be merged using the variable CNTSTUID

Missing code conventions

Data for a variable can be missing for different reasons. The data files use different coding to represent missing data. Reasons for missing data can be one of the following:

- Missing/blank – 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.
- No response/omit – The respondent had an opportunity to answer the question but did not respond. For derived variables, this can also indicate that data were incomplete for a component variable.
- Invalid – Used to indicate a questionnaire item was suppressed by country request or that an answer was not conforming to the expected valid response options. For a paper-based questionnaire, it is used when the respondent indicated more than one choice for an exclusive-choice question. For a computer-based questionnaire, it is used when the response was not in an acceptable range of responses, e.g., the response to a question asking for a percentage was greater than 100.
- Not applicable – A response was provided even though the response to an earlier question directed the respondent to skip that question, or the response could not be determined due to a printing problem or torn booklet. In the questionnaire data, it is also used to indicate a response missing by design (i.e. the respondent was never given the opportunity to answer this question).
- Valid skip – The question was not answered because a response to an earlier question directed the respondent to skip the question.

CODEBOOKS FOR THE PISA 2018 PUBLIC USE DATA FILES

Included with the PISA 2018 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/*).

The codebooks contain variable names, variable labels, values and value labels. Other metadata are 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 the file CY07_MSU_CODEBOOK.XLSX in separate worksheets that correspond with the eight public-use data files described above.

DATA COMPENDIA TABLES

Using the public use files as the source data, the compendia are sets of tables that provide percentages for both cognitive and background items. The compendia support public use file users so that they can gain knowledge of the contents of the data files and use the compendia results as reference for their quality control procedures for reading the data. The compendia are available on the OECD website (*http://www.oecd.org/pisa/*).

Questionnaire compendia provide frequency distributions for the categorical variables collected through the questionnaires. Cognitive compendia provide the distribution of student responses for each test item. Results are provided in Excel format, separately for background questions and test items, and are further broken out by type of questionnaire and by domain, and by gender for cognitive compendia). Each Excel file contains multiple worksheets, with each worksheet corresponding to a single variable. The first worksheet in each file is a table of contents that contains a hyperlink to each variable so users can see at a glance which variables are available and can click to go directly to the desired data.

Separate tables are provided with percentage and percentile data for continuous background variables across all questionnaires.

All statistics are calculated using weighted data, with their corresponding standard errors taking into account sampling and measurement uncertainty. The OECD average is created from the 37 current OECD member countries.

DATA ANALYSIS AND SOFTWARE TOOLS

Standard analytical packages for the social sciences and educational research do not readily recognise or support handling 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 analysis outputs: the PISA Data Explorer (PDX) and the IEA's IDB Analyzer. Each of these two software tools addresses a slightly different set of needs. While the PDX is a web-based application that allows relatively easy calculation of means, totals and proportions, the IEA's IDB Analyzer, used in conjunction with the 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 Data Explorer (PDX)

The PDX is a web-based application that allows the user to query an OECD hosted, secure, PISA International Database via a web browser. In addition to the PISA 2018 micro-data, the PDX database contains micro-data from previous cycle PISA international that was released in public use files. The PDX is available on the OECD website (*http://www.oecd.org/pisa/*). Using the PDX, the user can navigate, analyse, and produce report quality tables and graphics.

The database underlying the PDX is populated using the public use files that include more than 2.4 million unique student records across seven PISA cycles. Over 5,000 variables across seven assessment cycles and more than 100 countries and adjudicated subregions are available for analysis. Because certain variables that are included in the public use file (PUF) for secondary analysis are not informative as part of the PDX, they are not included in the PDX database. The majority of variables included only in the PUF relate to the individual cognitive item scores and timing and process information.

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 in Microsoft Word, Microsoft Excel and HTML formats.

In the PISA Data Explorer, the OECD average is created from the 37 current OECD member countries. The same 37 countries are used to create the OECD average for all previous PISA cycles of data.

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, submit the request to a central server where processing takes place, and then receive and display back the results in a user friendly and publication ready 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 used 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, within and across cycles, while implementing necessary adjustments that take into account the sample and test design, as well as adjustment for multiple comparisons. Significance test results can be displayed in table or in graphic format.

Table results can be easily exported and 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 and used in reports and presentations.

The web application is compatible with many widely used browsers including Internet Explorer 10 and higher, Firefox 3.0 and higher, Google Chrome, and Safari. Target screen resolution is 1024x768. To fully take advantage of the PDX, users should enable JavaScript and pop-ups in

IEA'S INTERNATIONAL DATABASE ANALYZER

The IEA 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 OECD, such as PISA. Originally designed for IEA's international large-scale assessments, it is also capable of working with national assessments such as the US 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. It generates SPSS or SAS syntax that 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. While anyone can use the software for free, users do not have ownership of the software itself or its components, including the SPSS and SAS macros, and users are only authorised to use the SPSS and SAS macros in combination with the IDB Analyzer, unless explicitly authorised by the IEA. The software and license expire at the end of each calendar year, when the user will again have to download and reinstall the most current version of the software, and agree to the new license. 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 regression coefficients (linear and logistic) 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. The following analyses can be performed with the analysis module:

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

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

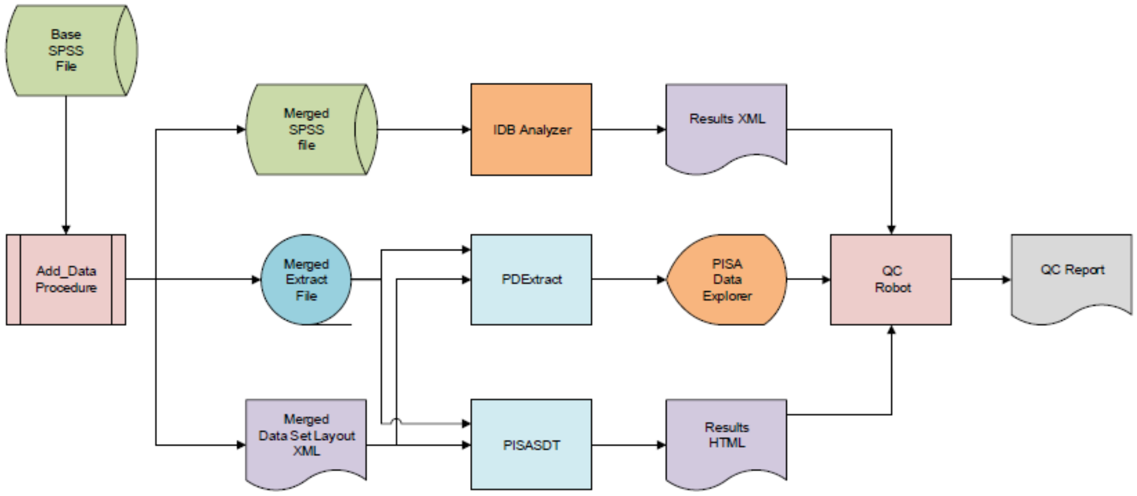
When calculating these statistics, the IDB Analyzer has the capability of using any continuous or categorical variable in the database or make use of scores in the form of plausible values. When using plausible values, the IDB Analyzer generates SPSS or SAS code that takes into account the multiple imputation methodology in the calculation of the variance for statistics, as it applies to the corresponding study.

The IDB Analyzer is configured to recognize the data structure from current PISA data (using 10 plausible values), as well as legacy PISA data (using 5 plausible values) and analyse the data accordingly.

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 19.1 below. This process was applied separately to the data from each country.

Figure 19.1 : PISA database population and quality control



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

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 is 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.