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 IDB Analyzer was configured to work with PISA data and can be downloaded at *http://www.iea.nl/our-data*.

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. Each national database includes approximately 3 000 common variables for student cognitive and background questionnaire assessments and approximately 600 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/*.

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 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 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 background questionnaire (BQ) or process variables that are available (e.g. country and language), especially detailed free-text entry items
3. all national adaptations and extensions in the BQ
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 five public use data files: the student questionnaire data file (which also includes estimates of student performance and parent-questionnaire data), the school questionnaire data file, the teacher questionnaire data file, the cognitive item data file and a file with questionnaire timing data. These files include countries/economies/subregions that fully met adjudication criteria. An additional data file contains the data for countries with adjudication issues.

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

- **Student questionnaire data file (PUF_COMBINED_CMB_STU_QQQ.zip):** This file includes ID variables, all student questionnaire data (from the Student Background Questionnaire, Educational Career Questionnaire, and Information and Communication Technology Questionnaire), parent-questionnaire data, student and parent-questionnaire 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.
- **Cognitive Item data file (PUF_COMBINED_CMB_STU_COG.zip):** The cognitive data file includes ID variables, raw and coded items, computer-based assessment (CBA) item log data (total time and number of actions), as well as some additional CBA cognitive new science information.

- **Questionnaire timing data file (PUF_COMBINED_CMB_STU_QTM.zip):** The questionnaire timing data file includes CBA questionnaire log data (i.e., total time on a unit/screen).
- **Additional data files for Albania, Argentina, Kazakhstan and Malaysia (PUF_COMBINED_CM2_STU_QQQ_COG_QTM_SCH_TCH.zip):** These files include all data for Argentina, Kazakhstan and Malaysia, and student questionnaire data for Albania. Due to issues identified during data adjudication, caution is required when these data. For further information, see *Annex A4* of *PISA 2015 Results (Volume I): Excellence and Equity in Education* (OECD, 2016).

Data for student questionnaire items ST016 and ST038 are made available in the *PISA 2015 Results Volume III*, published in April 2017. Financial literacy datasets are available in the *PISA 2015 Results Volume IV*, published in May 2017. Collaborative problem solving datasets are available in the *PISA 2015 Results Volume V*, published in November 2017.

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 *SENWT* 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 makes the population of each country to be 5 000 to ensure an equal contribution by each of the countries in the analysis. 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 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 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 five MISSING categories:

1. 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.
2. No response/omit – The respondent had an opportunity to answer the question but did not respond.
3. Invalid – Used to indicate a questionnaire item was suppressed by country request or that an answer was not conforming to the expected response. For a paper-based questionnaire, the respondent indicated more than one choice for an exclusive-choice question. For a computer-based questionnaire, 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.
4. Not applicable – A response was provided even though the response to an earlier question should have 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 missing by design (i.e. the respondent was never given the opportunity to see this question).
5. Valid skip – The question was not answered because a response to an earlier question directed the respondent to skip the question. This code is assigned by Core 3 during data processing.

CODEBOOKS FOR THE PISA 2015 PUBLIC USE DATA FILES

Included with the PISA 2015 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 information is displayed with 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 five 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. Timing – Student questionnaire timing data
5. Teacher – Teacher questionnaire data.

Codebooks for the additional files for Albania, Argentina, Kazakhstan and Malaysia are contained in a similar set of worksheets in the file **Codebook_CMS.xlsx**.

DATA COMPENDIA TABLES

Using the public use files as the source data, the compendia are sets of tables that provide categorical 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 so that they are performing public use file analyses correctly. The compendia are available on the OECD website (*http://www.oecd.org/pisa/*).

Questionnaire compendia provide the distribution of students according to the 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 tables). 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.

For each questionnaire (EC, ICT, Parent, School, and Student), the percentage of responses in each category are provided in the Excel files with "overall" in the name. Average scale scores corresponding to each category are provided in the files identified by the domains "math", "read", and "scie". The file "**pisa_bq_continuous_overall_compendium.xls**" provides 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 35 current OECD member countries.

The nine Excel files for the cognitive data provide percentages in each response category for the test items. Results are provided separately for females, males, and overall (total) for each domain.

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 outputs: the PISA Data Explorer (PDX) and a micro-data analyser. 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 and publication-ready access to basic estimates 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 PISA 2015 micro-data, the PDX database contains previous cycle PISA international micro-data that was released in public use files. The PDX is available on the OECD website (*http://www.oecd.org/pisa/*) and provides access to a secure PISA database (protected by the OECD firewalls and security mechanisms) to navigate, analyse, and produce report quality tables and graphics.

The database underlying the PDX is populated using the public use files to import more than 2.4 million unique student records across six PISA cycles. About 5,000 variables across six assessment cycles and over 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 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.

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 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. Export formats currently available include MS Word, MS Excel, PDF and HTML.

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 7 and higher, Firefox 3.0 and higher, Google Chrome, and Safari. Target screen resolution is 1024x768. Users should enable JavaScript and pop-ups in their browsers and install Adobe Flash Player 9.0.115 or higher.

Import of trend data

The PISA trend data from 2000 to 2012 were imported into the PDX directly from a database that had been established earlier by the United States Department of Education to develop and support a Data Explorer for PISA and other international studies. These data were taken from all public use files that were available for those cycles and were updated with all subsequent releases of modified or additional data. This approach ensured that all calculated results were consistent with all available OECD reports.

An important outcome of this prior work was the establishment of a naming convention for all data variables to ensure that valid trend comparisons could be made across cycles, even though the variable names as used in the public use file data were not consistent across cycles. This naming convention was extended and applied to all of the 2015 variables in order to ensure continuity and comparability with previous cycles.

In the PISA Data Explorer, the OECD average is created from the 35 current OECD member countries. The same 35 countries are used to create the OECD average for all previous PISA cycles.

Trend comparison link error factors

Comparisons of performance between two assessments in each domain (e.g., a country's/economy's change in performance between PISA 2000 and PISA 2015 or the change in performance of a subgroup) are calculated using the link error factors shown in Table 19.1.

Table 19.1 Robust link error for comparisons of performance between PISA 2015 and previous assessments

Comparison	Mathematics	Reading	Science	Financial literacy
PISA 2000 to 2015		6.8044		
PISA 2003 to 2015	5.6080	5.3907		
PISA 2006 to 2015	3.5111	6.6064	4.4821	
PISA 2009 to 2015	3.7853	3.4301	4.5016	
PISA 2012 to 2015	3.5462	5.2535	3.9228	5.3309

Note: Comparisons between PISA 2015 scores and previous assessments can only be made when the subject first became a major domain. As a result, comparisons in mathematics performance between PISA 2015 and PISA 2000 are not possible, nor are comparisons in science performance between PISA 2015 and PISA 2000, or PISA 2003.

INTERNATIONAL DATABASE ANALYZER

The IEA International Database Analyzer (IDB Analyzer)[1] is an application developed by the IEA Data Processing and Research Center (IEA-DPC) in Hamburg, Germany, 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 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 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:

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).
8. Differences by Performance Groups: Computes the means on an analysis variable by subgroups defined by background variables and performance level. When there are two subgroups within a performance level, it computes significance testing of the difference between these two groups. Currently this functionality is only available with SPSS.

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.

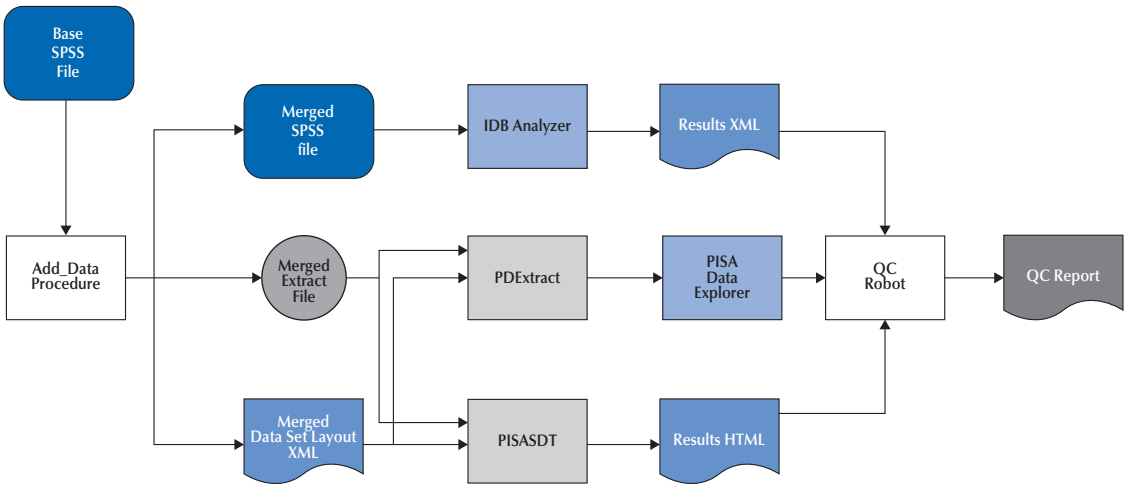
All procedures offered within the analysis module of the IDB Analyzer make use of appropriate sampling weights and standard errors of the statistics that are computed according to the variance estimation procedure required by the design as it applies to the corresponding study.

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.

Note

1. *http://www.iea.nl/our-data*.

Reference

OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, OECD Publishing, Paris, *http://dx.doi.org/10.1787/9789264266490-en*.