PISA 2022 Technical Report

OECD

22 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 from www.iea.nl.

Public-use files

The public-use files (PUF) contain response records from all participating countries/economies that are part of the approved PISA sample. Student-level files contain over 6000 variables that include responses to the background questionnaire and the cognitive assessments, as well as sampling weights, proficiency estimates and variables derived from responses to the background questionnaire. The student and teacher files contain over 1000 variables. The variables included in the PUF represent a common subset of the variables that were collected across all participating countries/economies and are available on the OECD website at *http://www.oecd.org/pisa/*.

Variables excluded or suppressed for some or all countries

The PUF include only a subset of the variables included in the individual country files. The PUF do not include any data collected using national adaptations and extensions. Rather, they include common data that were collected or derived across all countries. Additionally, variables were also excluded after 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 PUF are:

- direct, indirect, and operational identifiers for respondents;
- certain background questionnaire (BQ) or process variables such as free text entry responses and random numbers used by the SDS to determine routing;;
- all national adaptations and extensions in the BQ;
- original scale score values (theta) before standardisation to an international metric.

Countries were given the option of suppressing variables in the PUF. 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.

Data files

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

- **Student questionnaire data file:** This file includes ID variables, all student questionnaire response data, parent-questionnaire response data, student and parent background questionnaire scale and derived variables, plausible values for the core domains (Reading, Math, and Science), and overall and replicate student weights.

- **School questionnaire data file:** The school questionnaire data file includes ID variables, school questionnaire response data, school questionnaire scale and derived variables, and an overall school weight.
- **Teacher questionnaire data file:** The teacher questionnaire data file includes ID variables, teacher questionnaire response data, and teacher questionnaire scale and derived variables, and overall and replicate teacher weights.
- **Cognitive item data file[1]:** The cognitive data file includes ID variables, raw and coded item responses, item log data for the computer-based assessment (e.g., total time and number of actions) for the core domains (Mathematics, Reading, Science).
- **Creative Thinking cognitive data file:** The cognitive data file includes ID variables, Creative Thinking raw and coded item responses, computer-based assessment (CBA) item log data (total time and number of actions); and Creative Thinking plausible values including the Maths, Reading, and Science plausible values that were created as part of the population model with the Creative Thinking cognitive data.
- **Financial Literacy student questionnaire data file[2]:** This file includes ID variables, all student questionnaire response data, parent-questionnaire response data, student and parent background questionnaire scale and derived variables, plausible values for the domains assessed (Financial Literacy, Reading, and Maths), and overall and replicate student weights for the optional financial literacy sample.
- **Financial Literacy cognitive item data file[12]:** The cognitive data file includes ID variables, raw and coded item responses, item log data for the computer-based assessment (e.g., total time and number of actions) for the domains assessed in the Financial Literacy sample (Financial Literacy, Maths, Reading).
- **Questionnaire timing data file:** The questionnaire timing data file includes CBA questionnaire log data (i.e., total time on a unit/screen).

The Creative Thinking datasets *and* Financial Literacy datasets are scheduled to be published in 2024.

Variables used in sampling, weighting and merging

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

The variables W_FSTUWT and W_FSTURWT1 - W_FSTURWT80 represent the full student sampling weight, and the 80 replicate weights used for estimation of sampling variance.

The variable *SENWT* is a normalised weight variable typically used for analyses of student performance across a group of countries where contributions from each of the countries in the analysis is desired to be equal regardless of their population or sample size. The senate weight adds to a constant of 5 000 across all cases within each country/economy in the file. This weight adds to 5000 within each country/economy only when there is no missing data for the variable of interest. The relative contribution of each country/economy is affected by the incidence 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 school ID number, making it unique across all countries.

Codebooks for the PISA 2022 public-use data files

Included with the PISA 2022 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 categorical variables from the cognitive and background questionnaire variables, as well as those that have been derived and/or added during data processing. The codebooks are available from the OECD website (https://www.oecd.org/pisa/data/).

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 the range of valid values (minimum and maximum) for non-categorical numeric variables.

Codebooks for the main files are contained in separate worksheets within the file made available at the OECD website. Each worksheet corresponds to one of the eight public-use data files described above.

Data compendia tables

Using the PUF as the source data, the compendia are sets of summary 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 to confirm that they are performing analyses on the PUF 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 with 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 items). 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 including in the compendia are calculated using weighted data and are presented with their corresponding standard error that take into account both the sampling and measurement uncertainty. The OECD average is created as the simple average of the 38 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 outputs: The PISA Data Explorer (PDX) and the IEA's International Database Analyzer (IDB Analyzer). Each of these two software tools address 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 the PISA 2022 data, the PDX database contains data from previous cycles of PISA. The PDX is available on the OECD website (https://pisadataexplorer.oecd.org/ide/idepisa/) and provides access to a secure PISA database that is protected by the OECD firewalls and security mechanisms. The PDX allows the user to navigate, analyse, and produce report quality tables and graphics.

The database underlying the PDX is populated using the PUF to import more than 3.5 million unique student records across eight PISA cycles. About 8,700 variables across eight assessment cycles and over 100 countries, economies, 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 in the PUF but not the PDX relate to the individual cognitive item responses 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.

In the PISA Data Explorer, the International Average (OECD) includes all OECD member countries for which data are available for the corresponding subject and year (38 OECD Member countries as of PISA 2022). Depending on data availability, the countries contributing to this average might vary by cycle and subject.

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 variables selected, 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 Microsoft Edge, Firefox, Google Chrome, and Safari.

Import of trend data

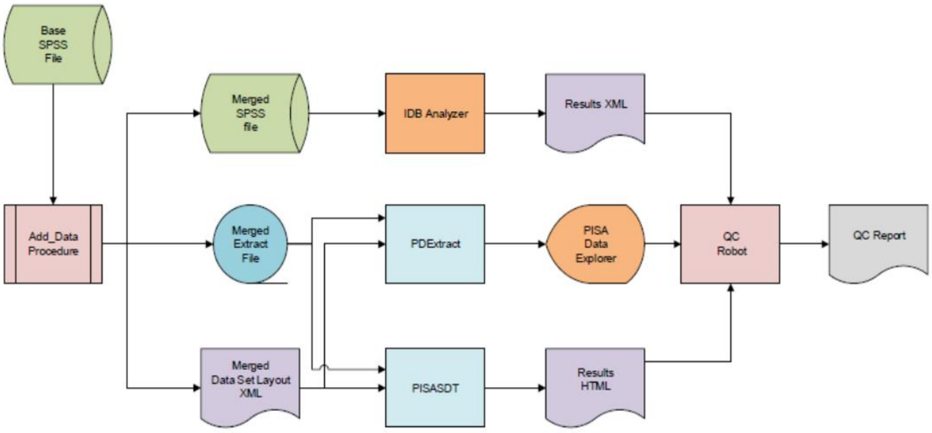
The PISA trend data from 2000 to 2012 were imported into the PDX directly from a database that had been established earlier by the United States Department of Education to develop and support a Data Explorer for PISA and other international studies. These data were taken from all PUF that were available 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 variables in subsequent PISA cycles (2015, 2018, and 2022) in order to ensure continuity and comparability with previous cycles.

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

Figure 22.1. 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 has produced and is available upon request to the OECD Secretariat [3]. 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 International Database Analyzer (IDB Analyzer) is an application developed by the International Association for the Evaluation of Educational Achievement (IEA) that can be used to analyse data from most major large-scale assessment surveys, including those conducted by the OECD, such as PISA. Originally designed for international large-scale assessments, it is also capable of working with national assessments such as the United States National Assessment of Educational Progress (NAEP).

The IDB Analyzer creates SPSS, SAS, or R syntax that can be used to perform analysis with these international databases. The syntax considers information from the sampling design in the computation of sampling variance and handles the multiple plausible value imputations. 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, SAS, or R macros, and users are only authorised to use the SPSS, SAS, and R 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:

- 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 as well as the computation of group mean and percent differences between groups taking into account sample dependency.
- 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, SAS, or R 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.

Notes

1. After the analysis phased completed, it was determined that 4 students in Iceland's grade-based sample were analysed along with Iceland's main sample data. As a result, the public use data for Iceland excludes these 4 students, yet they are still included in PISA 2022 technical report tables where Iceland data are referenced.

2. For Financial Literacy, only a subset of participants for Canada and Belgium received the Financial Literacy assessment and it is not a nationally representative sample. Only the Belgium Flemish community as well as the Canadian provinces British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario and Prince Edward Island, participated in Financial Literacy for PISA 2022.

3. via email to EDU.Pisa@oecd.org

Note by the Republic of Türkiye
The information in this document with reference to “Cyprus” relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Türkiye recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Türkiye shall preserve its position concerning the “Cyprus issue”.

Note by all the European Union Member States of the OECD and the European Union
The Republic of Cyprus is recognised by all members of the United Nations with the exception of Türkiye. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.