18

International Database

FILES IN THE DATABASE

The PISA 2009 international database consists of five data files: three with student responses, one with school responses and one with parent responses. All are provided in text (or ASCII format) with the corresponding SAS® and SPSS® control files.

Student files

The student performance and questionnaire data file (filename: INT_STQ09_Dec10.txt; available at *http://pisa2009.acer.edu.au/*) contains, for each student who participated in the assessment, the following information:

- identification variables for the country, school and student;
- the student responses to the four questionnaires, i.e. the student questionnaire, reading for school (RFS) questionnaire, the international option information communication technology (ICT) questionnaire and education career (EC) questionnaire;
- the indices derived from each student's responses to the original questions in the questionnaires;
- the students' performance scores in mathematics, reading, science, and the five subscales of reading (five plausible values for each domain);
- the student weight variable and 80 Fay's replicates for the computation of the sampling variance estimates;
- weight factor to compute normalised (replicate) weights for countries' multi-level analysis;
- three sampling related variables: the randomised final variance stratum, the final variance unit and the original explicit strata, mostly labelled by country;
- test language variable from the cognitive test; and
- database version with the date of the release.

Two sets of indices are provided in the student questionnaire files. The first set is based on a transformation of one variable or it is based on a combination of information gathered from two or more variables. Twenty-seven indices of the first type are included in the database. The second set is the result of a Rasch scaling and consists of weighted likelihood estimate indices. Twenty-two indices from the student questionnaire and seven indices from the information communication technology questionnaire are included in the database from this second type. The PISA index of economic, social and cultural status (ESCS) is derived as factor scores from a principal component analysis and is also included in the database. For a full description of the indices, see Chapter 16.

For each domain, reading, mathematics and science, and for each scale in reading, i.e. *access and retrieve, integrate and interpret, reflect and evaluate, continuous text and non-continuous text*, a set of five plausible values transformed to the PISA scale are provided.

It is important to note that three aspect scales and two text format scales are based on the same test items. As such, it is inappropriate to jointly analyse any of the three aspect scales with any of the two text format scales. For example, it would not be meaningful to correlate or otherwise compare performance on the *access and retrieve* scale, with performance on the *continuous text* scale as some of the items are included in both of these two scales.

The metrics of the various scales are established so that in the year that the scale is first established the OECD students' mean score is 500 and the pooled OECD standard deviation is 100.[1] The reading scale was established in 2000, the mathematics scale in 2003 and the science scale in 2006. When establishing the scale, the data is weighted to ensure that each OECD country is given equal weight.

Plausible values for reading were mapped to the PISA 2000 scale, plausible values for mathematics were mapped to the PISA 2003 scale and plausible values for science were mapped to the PISA 2006 scale. See Chapter 12 for details of these mappings.

The variable *W_FSTUWT* is the final student weight. The sum of the weights constitutes an estimate of the size of the target population. When analysing weighted data at the international level, large countries have a greater contribution to the results than small countries. This weighting is used for the OECD total in the tables of the international report for the first results from PISA 2009 (OECD, 2010b). To weight all countries equally for a summary statistic, the OECD average is computed and reported. The OECD average is computed as follows. First, the statistic of interest is computed for each OECD country using the final student weights. Second, the mean of the country statistics is computed and reported as the OECD average.[2]

For a full description of the weighting methodology and the calculation of the weights, see Chapter 8. How to use weights in analysis of the database is described in detail in the *PISA Data Analysis Manual* for SPSS® or SAS® users (OECD, 2009),[3] which is available at *www.pisa.oecd.org*. The data analysis manual also explains the theory behind sampling, plausible values and replication methodology and how to compute standard errors in case of two-stage, stratified sampling designs.

Two files with student cognitive data are available. One file contains single digit and original responses (filename: INT_Cog09_TD_Dec10.txt; available at *http://pisa2009.acer.edu.au/)*. The second file contains scored responses (filename: INT_Cogn09_S_Dec10.txt; available at *http://pisa2009.acer.edu.au/)*.

For each student who participated in the assessment, the following information is available:

- Identification variables for the country, school and student.
- Test booklet identification.
- The student responses to the cognitive items. When the original responses consist of multiple digits (complex multiple choice or open ended items), the multiple digits were recoded into single digit variables for use in scaling software). A “T” was added to the end of the recoded single digit variable names. The original response variables have been added at the end of the single digit, unscored file (with an “R“ at the end of the variable name see further below). For the double-digit variables (M155Q02, M155Q03, M462Q01, S131Q02, S131Q04, S269Q03, S438Q03) a “D” was added to the end of the recoded single-digit variable.
- Test language.
- Database version with the date of the release.

The PISA items are organised into units. Each unit consists of a stimulus (consisting of a piece of text or related texts, pictures or graphs) followed by one or more questions. A unit is identified by a short label and by a long label. The units’ short labels consist of four characters and form the first part of the variable names in the data files. The first character is R, M or S for reading, mathematics or science, respectively. The next three characters indicate the unit within the domain.

For example, M155 is a mathematics unit. The item names (usually seven or eight digits) represent questions within a unit and are used as variable names (in the current example the item names within the unit are M155Q01, M155Q02D, M155Q03D and M155Q04T). Thus items within a unit have the same initial four characters plus a question number.

Responses that needed to be recoded into single digit variables have a “T“ or “D“ at the end of the variable name. The original multiple digit responses have been added to the end of the single digit and original responses file (*filename: INT_Cogn09_TD_Dec10.txt*) with an “R“ at the end of the variable name (for example, the variable M155Q02D is a recoded item with the corresponding original responses in M155Q02R at the end of the file).

The full variable label indicates the domain the unit belongs to, the PISA cycle in which the item was first used, the full name of the unit and the question number. For example, the variable label for M155Q01 is “MATH - P2000 POPULATION PYRAMIDS (Q01)“.

The scored data file (INT_Cogn09_S_Dec10.txt) only includes one single digit variable per item with scores instead of response categories.

In both files, the cognitive items are sorted by domain and alphabetically by item name within domain. This means that the mathematics items appear at the beginning of the file, followed by the reading items and then the science items. Within domains, units with smaller numeric identification appear before those with larger identification, and within each unit, the first question will precede the second, and so on.

School file

The school questionnaire data file (filename: INT_SCQ09_Dec10.txt; available at *http://pisa2009.acer.edu.au/*) contains the following information for each school that participated in the assessment:

- the identification variables for the country and school;
- the school responses on the school questionnaire;
- the school indices derived from the original questions in the school questionnaire;
- the school weight;

- explicit strata with national labels; and
- database version with the date of the release.

The school file contains the original variables collected through the school context questionnaire. In addition, two types of indices are provided in the school questionnaire files. The first set is based on a transformation of one variable or on a combination of two or more variables. The database includes 10 indices from this first type. The second set is the result of a Rasch scaling and consists of weighted likelihood estimate indices. Nine indices are included in the database from this second type. For a full description of the indices and how to interpret them see Chapter 16. The school weight (*W_FSCHWT*) is the trimmed school-base weight adjusted for non-response (see also Chapter 8).

Although the student samples were drawn from within a sample of schools, the school sample was designed to optimise the resulting sample of students, rather than to give an optimal sample of schools. For this reason, it is always preferable to analyse the school-level variables as attributes of students, rather than as elements in their own right (Gonzalez and Kennedy, 2003).

Following this recommendation one would not estimate the percentages of private schools versus public schools, for example, but rather the percentages of students attending a private school or public schools. From a practical point of view, this means that the school data should be merged with the student data file prior to analysis.

For general information about analysis of the data, see the *PISA Data Analysis Manual* for SPSS® or SAS® users (OECD, 2009),[4] also available at *www.pisa.oecd.org*. Chapter 10 of the data analysis manual describes analysis with school level variables. Chapter 15 is about multi-level analysis using PISA data.

Parent file

The parent questionnaire file (filename: INT_PAQ09_Dec10.txt, available at http://pisa2009.acer.edu.au/) contains the following information:

- identification variables for the country, school and student;
- the parents' responses on the parent questionnaire;
- the parent indices derived from the original questions in the parent questionnaire; and
- the database version with the date of the release.

The parent file contains the original variables collected through the parent context questionnaire as a national option instrument. In addition, two types of indices are provided in the parent questionnaire file. The first set is based on a transformation of one variable or on a combination of two or more variables. The database includes three indices from this first type. The second set is the result of a Rasch scaling and consists of weighted likelihood estimate indices. Six indices are included in the database from this second type. For a detailed description of the indices see Chapter 16.

Due to the high parent non-response in most countries, caution is needed when analysing this data. Non-response is unlikely to be random. When using the final student weights from the student file, the weights of valid students in the analysis do not sum up to the population size of parents of PISA eligible students. A weight adjustment is not provided in the database.

RECORDS IN THE DATABASE

Records included in the database

Student and parent files

- All PISA students who attended test (assessment) sessions.
- PISA students who only attended the questionnaire session are included if they provided at least one response to the student questionnaire and the father's or the mother's occupation is known from the student or the parent questionnaire.

School file

- All participating schools – that is, any school where at least 25% of the sampled eligible, non-excluded students were assessed – have a record in the school-level international database, regardless of whether the school returned the school questionnaire.

Records excluded from the database

Student and parent file

- Additional data collected by countries as part of national or international options.
- Sampled students who were reported as not eligible, students who were no longer at school, students who were excluded for physical, mental or linguistic reasons, and students who were absent on the testing day.
- Students who refused to participate in the assessment sessions.
- Students from schools where less than 25% of the sampled and eligible, non-excluded students participated.

School file

- Additional data collected by countries as part of national or international options.
- Schools where fewer than 25% of the sampled eligible, non-excluded students participated in the testing sessions.

REPRESENTING MISSING DATA

The coding of the data distinguishes between four different types of missing data:

- Item level non-response: 9 for a one-digit variable, 99 for a two-digit variable, 999 for a three-digit variable, and so on. Missing codes are shown in the codebooks. This missing code is used if the student or school principal was expected to answer a question, but no response was actually provided.
- Multiple or invalid responses: 8 for a one-digit variable, 98 for a two-digit variable, 998 for a three-digit variable, and so on. For the multiple-choice items code 8 is used when the student selected more than one alternative answer.
- Not-administered: 7 for a one-digit variable, 97 for a two-digit variables, 997 for a three-digit variable, and so on. Generally this code is used for cognitive and questionnaire items that were not administered to the students and for items that were deleted after assessment because of misprints or translation errors.
- Not reached items: all consecutive missing values clustered at the end of test session were replaced by the non-reached code, "r", except for the first value of the missing series, which is coded as item level non-response.

HOW ARE STUDENTS AND SCHOOLS IDENTIFIED?

The student identification from the student and parent files consists of three variables, which together form a unique identifier for each student:

- a country identification variable labelled *COUNTRY* – the country codes used in PISA are the ISO numerical three-digit country codes (*http://unstats.un.org/unsd/methods/m49/m49alpha.htm*);
- a school identification variable labelled *SCHOOLID*; and
- a student identification variable labelled *STIDSTD*.

A fourth variable has been included to differentiate adjudicated sub-national entities within countries. This variable (SUBNATIO) is used for three countries as follows:

- ***Belgium.*** The value "05601" is assigned to the Flemish region and "05600" to the French and German regions of Belgium.
- ***Spain.*** The value "72401" is assigned to Andalusia, "72402" to Aragon, "72403" to Asturias, "72404" to "Balearic Islands", "72405" to Canary Islands,"72406" to Cantabria, "72407" to Castile and Leon, "72409" to Catalonia, "72411" to Galicia, "72412" to La Rioja, "72413" to Madrid, "72414" to Murcia, "72415" to Navarre, "72416" to the Basque Country, "72418" to Ceuta and Melilla, and "72499" to the rest of Spain.
- ***United Kingdom.*** The value "82600" is assigned to England, Northern Ireland and Wales and the value "82620" is assigned to Scotland.

A fifth variable is added to make the identification of countries more convenient. The variable CNT uses the ISO 3166-1 ALPHA-3 classification (*http://unstats.un.org/unsd/methods/m49/m49alpha.htm*), which is based on alphabetical characters rather than numeric characters (for example, for Sweden has COUNTRY=752 and CNT=SWE). It should be noted that for Shanghai the China numerical code (COUNTRY=156) was used along with a three letter code "QCN" (the three letter code for China is CHN).

A sixth variable (*STRATUM*) is also included to differentiate sampling strata. Value labels are provided in the control files to indicate the population defined by each stratum.[5]

The school identification consists of two variables, which together form a unique identifier for each school:

- The country identification variable labelled *COUNTRY*. The country codes used in PISA are the ISO numerical three-digit country codes.
- The school identification variable labelled *SCHOOLID*.

DRA DATABASE

For the 19 countries that participated in the PISA 2009 digital reading assessment, a separate database was prepared.

With the exception of Colombia and Spain, the number of cases included in the DRA database is the same as the number of cases in the PISA 2009 international database. Colombia and Spain chose to subsample schools from their large national school sample – see Chapter 4 for details on DRA sampling. The weight and replicate weight variables for these two countries have been adjusted in the DRA database to reflect this subsampling. For all other countries, the DRA weights and the pencil and paper weights are identical.

The PISA DRA international database consists of four data files: three with student responses and one with school responses. All are provided in text (or ASCII format) with the corresponding *SAS*® and *SPSS*® control files.

Student files

Student performance and questionnaire data file (filename: ERA_STQ09_ June11.txt; available at *http://pisa2009.acer.edu.au/*).

For each student all the variables that are included in the international database are also included in DRA data file. The following additional information is also included:

- The students' performance scores in DRA (five plausible values).
- DRA Language variable.
- DRA Test Form.

Two files with student cognitive data are available. One file contains single digit and original responses (filename: ERA_Cog09_TD_June11.txt; available at *http://pisa2009.acer.edu.au/*). The second file contains scored responses (filename: ERA_Cogn09_S_ June11.txt; available at *http://pisa2009.acer.edu.au/*).

Additional information included in the DRA cognitive files is as follows:

- Original and coded responses for DRA items.
- DRA Language variable.
- DRA Test Form.

School file

The school questionnaire data file (filename: ERA_SCQ09_ June11.txt; available at *http://pisa2009.acer.edu.au/*).

The DRA school file contains the same information as the international data file for the participating countries.

FURTHER INFORMATION

A full description on how to analyse the PISA database in accordance with the complex methodologies used to collect and process the data is provided in the *PISA Data Analysis Manual* (OECD, 2009),[6] available at *www.pisa.oecd.org*.

Notes

1. The list of OECD countries included in each cycle when the scales were established is included in Annex J.

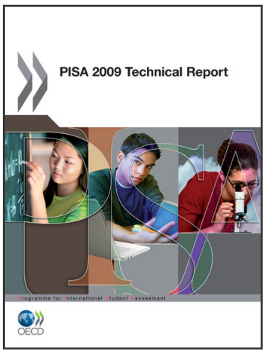
2. The definition of the OECD average has changed between PISA 2003 and PISA 2006. In previous cycles, the OECD average was based on a pooled, equally weighted database. To compute the OECD average, the data was weighted by an adjusted student weight variable that made the sum of the weights equal in all countries.

3. This publication is focused on PISA 2006, but the principles remain the same for PISA 2009.

4. This publication is focused on PISA 2006, but the principles remain the same for PISA 2009.

5. Note that not all participants permit the identification of all sampling strata in the database.

6. This publication is focused on PISA 2006, but the principles remain the same for PISA 2009.



From:

PISA 2009 Technical Report

Access the complete publication at:

https://doi.org/10.1787/9789264167872-en

Please cite this chapter as:

OECD (2012), “International Database”, in *PISA 2009 Technical Report*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264167872-19-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

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