19

International Databases

This chapter describes the three databases containing the PISA 2012 international data. The PISA 2012 International Database is described first. This description is followed by shorter descriptions of the computer-based assessment database and the financial literacy database, highlighting the differences between them and the PISA 2012 international database.

FILES IN THE DATABASE

The PISA 2012 international database consists of five data files: three with student responses, one with school responses and one with parent responses. All are provided in fixed width text (or ASCII) format with the corresponding SAS® and SPSS® control files.

Student files

The student performance and questionnaire data file (*filename: INT_STU12_DEC03.txt*, available at *www.oecd.org/pisa*) contains, for each student who participated in the assessment, the following information:

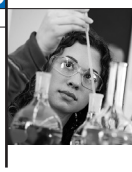
- Identification variables for the country, school and student.
- The student responses to the three questionnaires, *i.e.*, the student questionnaire, the information communication technology international option questionnaire, and the education career international option questionnaire.
- The indices derived from each student's responses to the original questions in the questionnaires.
- The students' performance scores in mathematics, reading, science, and the seven subscales of mathematics (five plausible values for each of these domains).
- The student weight variable and 80 Fay's replicates for the computation of the sampling variance estimates and a senate weight.
- 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 size. The senate weight makes the population of each country to be 1000 to ensure an equal contribution by each of the countries in the analysis. This weight is only applicable to the student performance scores (plausible values) that do not contain missing values. Its application to other variables might be compromised by its dependence on the patterns of missing data.
- Three sampling related variables: the randomised final variance stratum, the final variance unit and the original explicit strata, mostly labelled by country.
- A test language variable from the cognitive test.
- A database version identifier with the date of the release.

Three 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. Forty-nine 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-four indices from the Student Questionnaire and seven indices from the information communication technology questionnaire are included in the database from this second type. The third set is the result of applying an anchoring vignettes approach in the preparation of indices and a subsequent Rasch scaling of the indices, and consists of weighted likelihood estimate indices (see Chapter 16). Thirteen indices of this type are included in the database. 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, i.e. mathematics, reading and science, and for each subscale in mathematics, i.e. the four content categories of *change and relationships, quantity, space and shape, uncertainty and data,* as well as the three process categories of *employ, formulate* and *interpret,* a set of five plausible values (transformed to the PISA scale) are provided.

It is important to note that four content scales and three process scales are based on the same test items. As such, it is inappropriate to jointly analyse any of the four content scales with any of the three process scales. For example, it would not be meaningful to correlate or otherwise compare performance on the *space and shape* scale, with performance on the *employ* scale as some of the items are included in both of these two scales.

The metrics of the performance 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. The reading scale was established in



2000, the mathematics scale in 2003 and the science scale in 2006. When establishing the scale the data are weighted to ensure that each OECD adjudicated 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 these 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 2012 (OECD, 2014). 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.[1]

For a full description of the weighting methodology and the calculation of the weights, see Chapter 8. How to use weights in analyses of the database is described in detail in the *PISA Data Analysis Manual* for SPSS® or SAS® users (OECD, 2009a, 2009b),[2] which is available at *www.oecd.org/pisa/pisaproducts/*. The data analysis manual also explains the theory behind sampling, plausible values and replication methodology and how to compute standard errors in the case of two-stage, stratified sampling designs.

Two versions of the student cognitive files are available:

1) a version that contains single-digit and original responses (*filename: INT_COG12_DEC03.txt*, available at *www.oecd.org/pisa*) and
2) a version that contains scored responses (*filename: INT_COG12_S_DEC03.txt*, available at *www.oecd.org/pisa*)

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 (*PM155Q02, PM155Q03, PM462Q01, PS131Q02, PS131Q04, PS269Q03, PS438Q03*) 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 five characters and form the first part of the variable names in the data files. The first two characters are PR, PM or PS for reading, mathematics or science respectively in the pencil and paper component of the assessment. The next three characters indicate the unit within the domain. For example, PM155 is a paper and pencil mathematics unit. The item names (usually eight or nine digits) represent questions within a unit and are used as variable names (in the current example the variable names for items within the unit are *PM155Q01, PM155Q02D, PM155Q03D* and *PM155Q04T*). Thus items within a unit have the same initial five 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_COG_DEC03.txt*) with an "R" at the end of the variable name (for example, the variable *PM155Q02D* is a recoded item with the corresponding original responses in *PM155Q02R* 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 PM155Q01 is *'MATH - P2000 POPULATION PYRAMIDS (Q01)'*. The variable name for this item was *M155Q01* in PISA 2000-PISA 2009. Following the extended naming convention of PISA 2012, the variable label was modified to *PM155Q01* to reflect that it belongs to a set of items for the paper-based assessment.

The scored data file (*filename: INT_COG12_S_DEC03.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_SCQ12_DEC03.txt*, available at *www.oecd.org/pisa*) contains for each school that participated in the assessment, the following information:

- 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 and senate school weight.
- Explicit strata with national labels.
- 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 twenty one indices from this first type. The second set is the result of a Rasch scaling and consists of weighted likelihood estimate indices. Twelve 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 analyses of the data, see the *PISA Data Analysis Manual* for SPSS® or SAS® users (OECD, 2009a, 2009b),[3] also available at *www.oecd.org/pisa/pisaproducts/*. Chapter 10 of the *PISA Data Analysis Manual* describes analyses with school level variables. Chapter 15 is about multi-level analysis using PISA data.

Parent file

The parent questionnaire file (*filename: INT_PAQ12_DEC03.txt*, available at *www.oecd.org/pisa*) contains the following information:

- identification variables for the country, school and student;
- the parents' responses to the parent questionnaire;
- the parent indices derived from the original questions in the parent questionnaire; and
- 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 twenty three indices from this first type. The second set is the result of a Rasch scaling and consists of Weighted Likelihood Estimate indices. Five 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 these 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 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

In the student and parent files

- All PISA students who attended paper-based or computer-based test 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.

In the 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 of the answer options.
- 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 as a result of the balanced incomplete block test design used in PISA, 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 (code 9).

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 *CNT*. The values for this variable are drawn from the ISO 3166-1 ALPHA-3 classification (*http://unstats.un.org/unsd/methods/m49/m49alpha.htm*) used by the United Nations. Note that for several PISA 2012 participants the value for the *CNT* variable does not correspond to this classification system. This occurs for two

possible reasons. Firstly, where a National Centre represents only part of the country. The codes of this type are QCN for Shanghai representing part of China, QCY for part of Cyprus, QRS for the Perm region of the Russian Federation, QUA, QUB and QUC for the three states of the United States, Florida, Connecticut and Massachusetts correspondingly. Secondly, where the National Centre represented only part of the country in a previous cycle, and even though the full country is participating in PISA 2012, the *CNT* value has been preserved for consistency. The only participant's code of this type is ARE for the United Arab Emirates.

- A school identification variable labelled *SCHOOLID*.
- A student identification variable labelled *STIDSTD*.

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

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

Some additional identification variables are included in all data files of the database as follows:

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

- *Belgium*. The value "0560100" is assigned to the Flemish region and "0560000" to the French and German regions of Belgium
- *Spain*. The value "7240100" is assigned to Andalusia, "7240200" to Aragon, "7240300" to Asturias, "7240400" to "Balearic Islands", "7240600" to Cantabria, "7240700" to Castile and Leon, "7240900" to Catalonia, "7241000" to Extremadura, "7241100" to Galicia, "7241200" to La Rioja, "7241300" to Madrid, "7241400" to Murcia, "7241500" to Navarre, and "7241600" to Basque Country. The value "7240000" is assigned to the rest of the country.
- *United Kingdom*. The value "8260000" is assigned to England, Northern Ireland and Wales and the value "8262000" is assigned to Scotland.
- *Argentina*. The value of "0320100" is assigned to the Autonomous City of Buenos Aires. The value "0320000" is assigned to the rest of the country.
- *United Arab Emirates*. The value of "7840100" is assigned to Abu Dhabi and the value of "7840200" is assigned to Dubai. The value "7840000" is assigned to the rest of the country.
- *Perm region of the Russian Federation*. The value of "6430059" is assigned to Perm data which was collected separately from the Russian Federation data.
- *Florida, Connecticut and Massachusetts of the United States*. The value "8400100" is assigned to Florida, the value "8400200" is assigned to Connecticut and the value "8400300" is assigned to Massachusetts.

A variable *NC* is included to identify National Centres. The variable is based on the ISO numerical three-digit country codes (*http://unstats.un.org/unsd/methods/m49/m49alpha.htm*) with an addition of leading zero and a trailing double-digit code having value of "*00*" for the majority of the countries and a different value for those countries where administration of PISA 2012 was performed by separate National Centres or where not the whole country participated in the assessment. For example, the *NC* code for Australia is "*003600*". *NC* code of Shanghai representing a part of China is "*015601*" and *NC* code for Scotland where PISA 2012 was administered separately from the rest of the United Kingdom is "*082620*".

A variable *STRATUM* is also included to differentiate sampling strata. The variable is created as a concatenation of a 3-letter country code and a two-digit region identifier and two-digit original stratum identifier. Value labels are provided in the control files to indicate the population defined by each stratum.[4]

COMPUTER-BASED ASSESSMENT DATABASE

For the 44 countries that participated in PISA 2012 problem solving or problem solving and computer-based assessment of mathematics and reading literacy a separate database was prepared.

With the exception of Brazil, Italy and Spain the number of cases included in the computer-based assessment (CBA) database is the same as the number of cases in the PISA 2012 international database. Brazil, Italy and Spain chose to

subsample schools from their large national school sample — see Chapter 4 for details of CBA sampling. The weight and replicate weight variables for these three countries have been adjusted in the CBA database to reflect this subsampling. For all other countries, the CBA weights and paper-based weights are identical.

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

Student files

Student performance and questionnaire data file (*filename: CBA_STU12_MAR31.txt*, available at *www.oecd.org/pisa*).

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

- The students' performance scores in problem solving (five plausible values), computer-based mathematics and digital reading (five plausible values for each domain). It is necessary to note that 32 countries participated in both problem solving and computer-based assessment of literacy and 12 countries in problem-solving assessment only. Thus for 12 countries, performance scores in computer-based mathematics and digital reading contain missing values for all students.
- CBA Language variable.
- CBA Test Form.

Two versions of the student cognitive files are available:

1) A version that contains single-digit and original responses (*filename: CBA_COG12_MAR31.txt*, available at *www.oecd.org/pisa*) and
2) a version that contains scored responses (*filename: CBA_COG12_S_MAR31.txt*, available at *www.oecd.org/pisa*)

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

- Original and coded responses for CBA items
- CBA Language variable
- CBA Test Form

Parent file

The Parent Questionnaire data file (*filename: CBA_PAQ12_MAR31.txt*, available at *www.oecd.org/pisa*).

The CBA parent file contains the same information as the international data file for the participating countries.

School file

The School Questionnaire data file (*filename: CBA_SCQ12_MAR31.txt*, available at *www.oecd.org/pisa*).

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

FINANCIAL LITERACY DATABASE

For the 18 countries that participated in the optional PISA 2012 financial literacy assessment a separate database was prepared.

The sampling of students for the financial literacy assessment was performed in such a way that there was no overlap with the main international database. Thus students reported in this database have no common records with the students included in international database – see Chapter 4 for details of the financial literacy sampling. The weight and replicate weight variables were calculated for each country along with the weights for students in the international database with the only exception of Spain where a special file was created for financial literacy students.

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

Student files

Student performance and questionnaire data file (*filename: FIN_STU12_MAR31.txt*, available at *www.oecd.org/pisa*)

For each student who participated in financial literacy assessment all the variables that are included in the international database are also included in financial literacy data file. The following variables are included in addition to (or instead of) the variables in the international database:

- Instead of performance scores in mathematics, reading and science reported in the international database three sets of plausible values calculated using only data from students who participated in the financial literacy assessment. Those sets are performance scores in mathematics, reading and financial literacy (five plausible values for each domain).
- Sixteen questionnaire items added to the Student Questionnaire for the countries that participated in the financial literacy assessment.

Two versions of the student cognitive files are available:

- a version that contains single-digit and original responses (*filename: FIN_COG12_MAR31.txt*, available at *www.oecd.org/pisa*) and
- a version that contains scored responses (*filename: FIN_COG12_S_MAR31.txt*, available at *www.oecd.org/pisa*)

Additional information included in the financial literacy cognitive files is as following:

- Original and coded responses for financial literacy items.

Parent file

The Parent Questionnaire data file (*filename: FIN_PAQ12_MAR31.txt*, available at *www.oecd.org/pisa*)

The financial literacy parent file contains the same information as the international data file for the participating countries.

School file

The School Questionnaire data file (*filename: FIN_SCQ12_MAR31.txt*, available at *www.oecd.org/pisa*)

The financial literacy 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),[5] available at *www.pisa.oecd.org*.

Notes

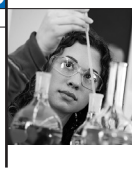
1. The definition of the OECD average has changed between PISA 2003 and PISA 2006. In PISA 2000 and 2003, the OECD average was based on a pooled, equally weighted database. To compute the OECD average the data was weighted by an adjusted student weight variable that made the sum of the weights equal in all countries.

2. This publication is focused on PISA 2006, but the principles remain the same for PISA 2012.

3. This publication is focused on PISA 2006, but the principles remain the same for PISA 2012.

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

5. This publication is focused on PISA 2006, but the principles remain the same for PISA 2012.



References

Gonzalez, E.J. and **A.M. Kennedy** (2003), *PIRLS 2001 User Guide for the International Database*, Boston College, Chestnut Hill.

OECD (2014), *PISA 2012 Results: What Students Know and Can Do - Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*, PISA, OECD Publishing, Paris.
http://dx.doi.org/10.1787/9789264208780-en

OECD (2009a), *PISA Data Analysis Manual: SPSS, Second Edition*, PISA, OECD Publishing, Paris.

OECD (2009b), *PISA Data Analysis Manual: SAS, Second Edition*, PISA, OECD Publishing, Paris.