Chapter 17: INTERNATIONAL DATA PRODUCTS

INTRODUCTION

After data processing and data analysis, public use data files and codebooks for PISA-D Strand C were delivered to OECD. These are available on the OECD website (https://www.oecd.org/pisa/pisa-for-development/database/).

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 participating countries. A subset of these variables were included in the public use data files, made available on the OECD website at (https://www.oecd.org/pisa/pisa-for-development/database/).

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 participating 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:

- direct, indirect and operational identifiers for respondents
- certain background questionnaire (BQ) variables, especially detailed free-text entry items
- all national adaptations and extensions in the BQ
- original scale score values (theta) before standardisation to an international metric.

As discussed in an earlier chapter, countries were given the option of suppressing variables in the public use files. Suppression of variables was approved when data presented a risk to respondent anonymity. Suppressed data are represented in the database by means of missing codes.

File names and content

There are three public use data files: the questionnaire data file, the questionnaire timing data file and the cognitive item data file.

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

- **Questionnaire (QQQ) data file:** This file includes ID variables, the Background Questionnaire responses, responses from the Person Most Knowledgeable (PMK) questionnaire, Background Questionnaire scale and derived variables, proficiency level values (reading and math), and full and replicate weights.
- **Questionnaire timing (TIM) data file:** This file includes participant questionnaire log data (i.e. total time on unit/screen).
- **Cognitive item (COG) data file:** The cognitive data file includes variables, raw and scored items, as well as log data (total time/time to first action/number of actions).

Variables used in sampling, weighting and merging

The variable *SPFWT0* contains the final full sample weight and the variables *SPFWT1* through *SPFWT30* contain the replicate weights used to calculate sampling variance. The variable *SENWT* is a normalised (senate) weight variable based on *SPFWT0* for analyses of participant 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 sum of the weights of each country be 5 000 to ensure an equal contribution by each of the countries in the analysis when countries are analysed combined. This weight is only applicable to the respondent 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 respondent data files can each be merged using the variable *CNTRSPID*. *CNTRSPID* is the combination of the three-digit country code and a randomised five-digit number, making it unique across all countries.

Missing code conventions

The data may include up to seven MISSING categories:

- Missing/blank (“.” in SAS; blank or “SYSMIS” in SPSS) – Used to indicate that the respondent was not presented the question according to the survey design or ended the questionnaire early and did not see the question.
- No response/omit (“.M” in SAS; “9/99/999/…” in SPSS) – Used to indicate the respondent had an opportunity to answer the question but did not respond. For derived variables, it is often used as an indicator for all different types of missing data.
- Invalid (“.I” in SAS; “8/98/998/…” in SPSS) – Used to indicate that the response was not appropriate or contradicted a prior response, e.g. the response to a question asking for a percentage was greater than 100.
- Not Reached (“.R” in SAS; “6/96/996/…” in SPSS) – Used in the cognitive scored variables to indicate that a respondent was unlikely to have seen the question and the response should be treated as such. This code is assigned during processing.
- Valid skip (“.V” in SAS; “5/95/995/…” in SPSS) – Used to indicate the question was not answered because a response to an earlier question directed the respondent to skip the

question. This code is assigned during data processing.

- Refused (“.F” in SAS; “4/4/994/…” in SPSS) – Used in the background questionnaire variables to indicate the respondent refused to answer the item.
- Don’t know (".D" in SAS; "3/93/993/…" in SPSS) - Used in the background questionnaire variables to indicate the respondent did not know the answer to the item.

Codebooks for the PISA-D Strand C Public Use Data Files

Included with the PISA-D Strand C 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 (https://www.oecd.org/pisa/pisa-for-development/database/).

The information is displayed with variable names and labels, values and value labels. Other metadata is 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 three separate worksheets (**CY1MDCI_Descriptives.xlsx**):

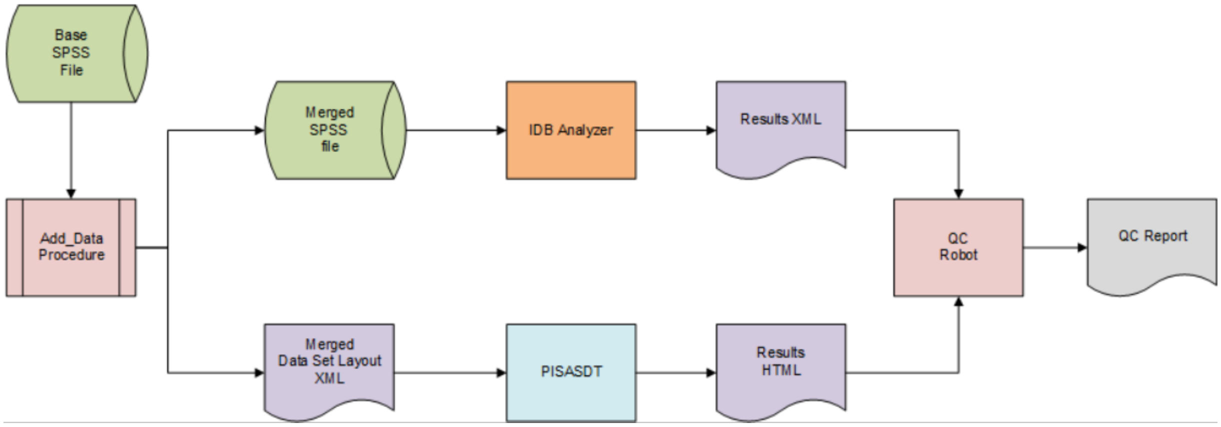
- Cognitive (COG) – respondent cognitive data for reading and mathematics
- Questionnaire (QQQ) – respondent questionnaire data including PMK questionnaire data
- Questionnaire Timing (TIM) – respondent questionnaire timing data.

QUALITY CHECK FOR THE PISA-D STRAND C DATA

The process to check the quality of the PISA-D Strand C database and confirm the results it produced is summarised in Figure 17.1. This process was applied separately to the data from each country.

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

Figure 17.1 Verification of Data Quality



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 PISASDT programme uses 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 set of benchmark levels. Each table in the analysis was a one-way tabulation of various statistics for each category of a given variable. The statistics include percentage by response category and percentages of the benchmark levels within each response category. Each statistic was accompanied by the standard error estimate, degrees of freedom, number of cases on which the statistic was 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.

The same analyses are conducted in SPSS using the IDB Analyzer to conduct the complex statistical calculations. The Analyzer reads data from the Base SPSS file, uses SPSS macros to calculate the desired statistics and writes the results on an XML file.

In the QC Robot procedure, the Results HTML documents from the PISASDT programme are compared to the corresponding Results XML document generated from the IDB Analyzer. The results of these comparisons are posted to the QC Report document where differences above specified criteria are flagged and subsequently examined.

Prior to the first execution of the procedure described above, the Analyzer and the PISASDT programmes were extensively calibrated with each other to ensure that the Merged SPSS and Merged Dataset Layout files were isomorphic and produced identical results for the statistics common to both programmes.