PISA 2022 Technical Report

12 Data Management Procedures

Introduction

In PISA, as in any international survey, standards and requirements for data collection guide the creation of an international database that allows for valid within-and-cross-country comparisons and inferences to be made. For both paper-based assessments (PBA) and computer-based assessments (CBA), these standards and requirements are developed with three major goals in mind: consistency, precision, and generalisability. To support these goals, data collection and management procedures are applied in a common and consistent way across all data to ensure data quality. As such, "data management" within the scope of the PISA survey refers to a collective set of procedures and tasks that each country performs to produce a verified, national database. With these procedures, national teams can avoid or, at the very least, minimise the potential for errors.

Although these international standards and requirements stipulate a collective agreement and mutual accountability among countries and contractors, PISA is an international study that includes countries with unique educational systems and cultural contexts. The PISA standards provide the opportunity for participants to adapt certain questions or procedures to suit local circumstances or add components specific to a particular national context. To handle these national adaptations, a series of consultations were conducted with the national representatives of participating countries to reflect country expectations in agreement with PISA 2022 technical standards. During these consultations, the data coding of the national adaptations to the instruments was discussed to ensure their recoding in a common international format. The guidelines for these data management consultations and recoding concerning national adaptations are described later in this chapter.

An important part of the data collection and management cycle is not only to control and adapt to the planned deviations from general standards and requirements, but also to control and account for the unplanned and/or unintended deviations that require further investigation by countries and contractors. Such deviations, at times, may compromise data quality and/or render data corrupt, or unusable. For example, it may be the case that implementing non-standard testing procedures might, in turn, affect test performance (e.g., session timing, the administration of test materials, and tools for support such as rulers and/or calculators). Sections of this chapter outline aspects of data management that are directed at controlling planned deviations, preventing errors, as well as identifying and correcting errors when they arise.

Given these complexities of large-scale assessment administration and the compressed PISA timeline, it remains an imperative task to record and standardise data procedures, as much as possible, with respect to the national and international standards of data management. These procedures are generalised to suit the individual cognitive test instruments and background questionnaire instruments used in each participating country. As a result, a suite of products is provided to countries to assist national teams in handling data management tasks in a standard way to prepare the national database and minimise errors. These products include a comprehensive data management manual, training sessions, as well as a range of other materials, including the data management software.

This chapter summarises these data management quality control processes and procedures and the collaborative efforts of contractors and countries to produce a final database for submission to the OECD.

Data management at the international and national level

Data management at the international level

To ensure compliance with the PISA technical standards, the following procedures were implemented by ETS Data Management to ensure data quality:

- Developed standards, guidelines, and recommendations for data management.
- Provided national teams with the data management software and developed data management manuals for modes of administration (PBA and CBA) as well as customized codebooks to support proper data capture.
- Facilitated data trainings and webinars and created hands-on, training resources (e.g., training exercises, lessons, and resource guides) for guided practice in building the national database and verifying data.
- Provided high-touch support for national team queries throughout the data management lifecycle.
- Enhanced data quality and verification procedures considering new context or situations during processing and cleaning data the international and national level.
- Prepared databases and reports for use by contractors, OECD, and the National Centres.
- Prepared interim and final data products (e.g., Data Explorer, compendia files) for dissemination to National Centres, the OECD, and, eventually, the public.

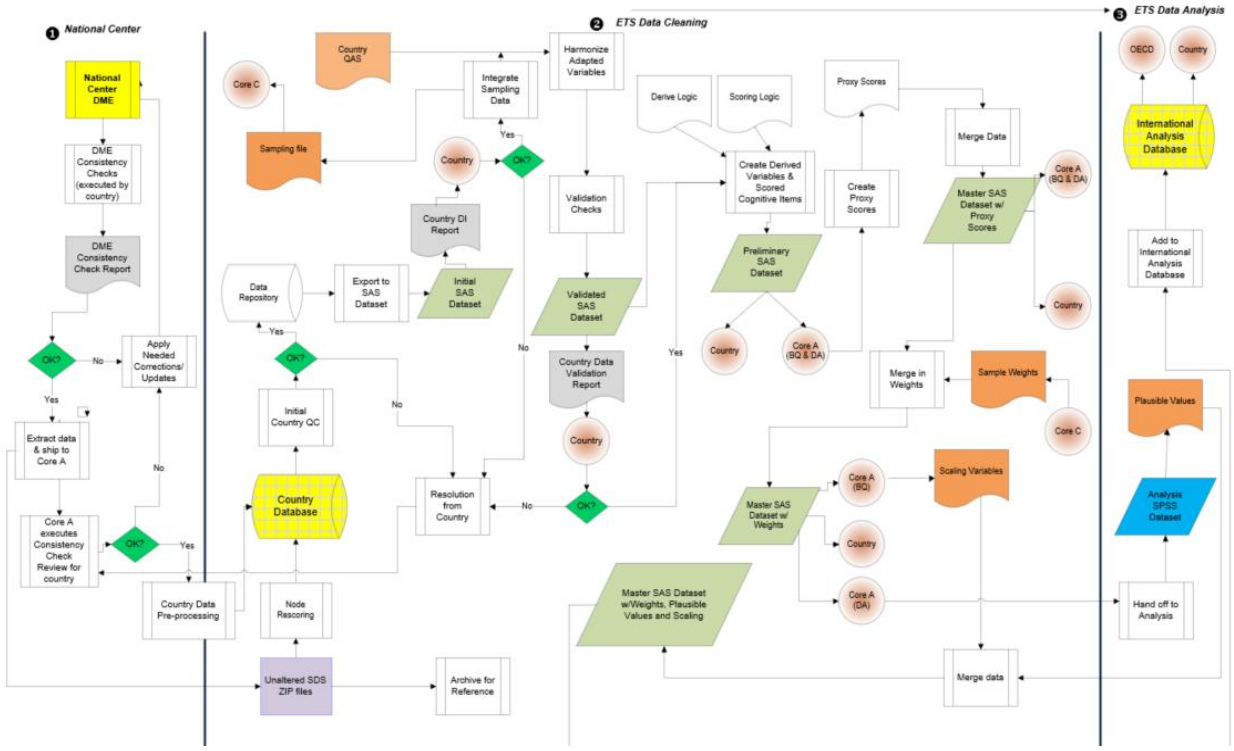
Ensuring compliance with technical standards also involved close collaboration with project partners. In PISA 2022, ETS Data Management worked closely with the all consortium members to ensure all data capture and quality procedures were accurately executed.

Data management at the national level

As the standards for data collection and submission involve a series of technical requirements and guidelines, each participating country appointed a National Project Manager (NPM) to organise the survey data collection and management at the National Centre. NPMs are responsible for ensuring that all required tasks, especially those relating to the production of a quality national database, are carried out on schedule and in accordance with the specified international standards and quality targets. The NPM is responsible for supervising, organising and delegating the required data management tasks at the national level. In addition, as these data management tasks require more technical skills of data analysis, NPMs were strongly recommended to appoint a National Data Manager (NDM) to complete all data related tasks on time and supervise support teams during data collection and data entry. These technical tasks for the NDM included, but were not limited to collaborating with ETS on template codebook adaptations; integration of data from the national PISA data systems (e.g. Student Delivery System, Open-Ended Coding System); manual capture of data after scoring for paper-based instruments; export/import of data required for coding (e.g. occupational coding); and data verification and validation with a series of consistency and validity checks.

To adhere to quality control standards, one of the most important tasks for National Centres concerned data entry and the execution of consistency checks from the primary data management software, the PISA Data Management Expert (DME). Figure 12.1 provides the workflow of the data management process for PISA 2022.

Figure 12.1. Overview of the data management process



The next section outlines the data management process as well as the application of additional quality assurance measures to ensure proper handling and generation of data. Additionally, more information is provided on the PISA 2022 DME as well as the phases of the data management cleaning and verification process.

The data management process and quality control

The collection of student, teacher, and school administrator responses on a computer platform into electronic data files provided a challenge and an opportunity for the accurate transcription of those responses as well as the collection of the associated process data, such as types of response actions and timing of those actions. It also requires a system that can accept and process these electronic data and their variety of formats as well as supports the manual entry of data from paper booklets and forms. To meet this challenge, ETS acquired a license for the use of the Data Management Expert (DME) software, which had previously proved successful in the collection and management of the data for the PISA 2015, PISA 2018, and PISA for Development large-scale surveys as well as the survey for adult skills (PIAAC) under a separate contract.

The DME is a high-performance.NET based, self-contained application that can be installed on most Windows operating systems (Windows XP or later), including Surface Pro and Mac, and does not require an internet connection to operate. It operates on a separate database file using SQLite constructed according to strict structural and relational specifications that define the data codebook. This codebook is a complete catalogue of all the data variables to be collected and managed, which are then arranged into well-defined datasets that correspond to the various instruments involved in the administration of the assessment. Before the datasets are created and ready for input processing, the application first validates the structure of the codebook to ensure the integrity of the database.

The first step in the data management process is to identify the different electronic and paper instruments, booklets, and forms that are to be collected and managed within each national centre and determine the variables to be collected from each instrument. These instruments and forms are then mapped into datasets, each containing their appropriate variables to form the international codebook, which will be the basis for every national codebook, whether the country is conducting the assessment on paper or computer. The international codebook is thoroughly checked, verified, and tested using marked up paper instruments as well as electronic data files that were created during testing of the various platforms.

The next step is the generation and testing of the national codebooks. Many of the variables used in the assessment and codebooks for PISA follow a systematic naming convention that provides additional information to the user. Table 12.1 describes the naming convention used in the codebooks and analysis.

Each national codebook is a copy of the international codebook where the datasets corresponding to national options implemented in the country are shown and the rest are hidden. For example, all codebooks for PBA countries will have the datasets corresponding to CBA instruments hidden from view and operation. In addition, the codebooks for CBA countries will have all adapted and national questions that were coded into the Questionnaire Adaptation Tool (QAT, described in Chapter 7) added to the appropriate datasets. The CBA codebooks are also tested using available test data obtained from the country's student delivery platform and the online questionnaire system. PBA countries, as well as CBA countries with the paper-based Parent Questionnaire, are given the option of providing national translations of all items in the paper instruments to be included in their national codebooks.

The codebook is delivered to each country as a national "template" file, containing the metadata the DME application uses to build the database file. The NDM must confirm that the template file will create an accurate codebook that supports the appropriate datasets for their national options. To verify nationally adapted variables and/or added national variables, CBA countries are then requested to also import available test data to confirm proper data capture. For PBA countries, variables must be added and adapted first to the questionnaire datasets, as there are no online QAT questionnaire data available for these countries. They are then required to test these adaptations and added variables with the manual entry of the questionnaire data to confirm that the variables are properly configured, in their correct sequence, and with their correct translations, when applicable. Similarly, CBA countries with the Parent Questionnaire option, a paper-based option, must also add and test their national adaptations to the corresponding dataset. After making all necessary modifications to and testing of their national codebook, every country is requested to send a copy of the codebook to Data Management so it can be reviewed for consistency and use in the Main Survey.

The DME application permits three levels of password-controlled access to the database – Administrator, Manager, and User. The Administrator level has complete access to all the database operations as well as the data tables and codebook-related tables. This level is reserved for Data Management. The Manager level is designated for the NDM in each country and includes the ability to make changes to the codebook, create and delete data tables and create User accounts and passwords, among other capabilities. The User level is assigned by the Manager for the purpose of creating clones of the project Master database to be used for manual data entry on multiple platforms. The DME application is designed to work in a distributed environment so that these individual clone databases can be easily merged into the master database.

For the PISA survey, there are three, recommended modes for input of data into the DME application: manual data entry, import from Excel or CSV file, and special import of extracted data from student delivery, sampling, and coding systems.

Manual data entry provides for the direct entry of data values into a targeted dataset through an interface that presents the description, format, and valid codes of each data element to be entered and validates each entered value. The type of forms that can be entered vary from a simple linear form, such as a questionnaire, to a series of booklets or forms that each contain a prescribed sequence of blocks of item

data, such as the cognitive booklets. The entry of the booklet/form number determines which variables are to be presented for entry and in what order. The manual entry mode is used primarily by PBA countries as well as those CBA countries when using the Parent Questionnaire option.

If a PBA country has its own data entry procedures in place, the data from these processes can be directly imported from Excel or CSV files where the first row/record contains the names of the variables whose data are in the corresponding columns. Again, all input data values are validated against the codebook and if any unexpected or out of range data values are found, the process stops. This import process has a corresponding export process to create files, typically Excel and CSV, from designated datasets. The two processes can be effectively used to move data into and out of the database. The export process for CSV files also produces syntax files for reading the exported data into SPSS or SAS so that separate analyses of the data can be performed with those applications.

The Export and Import functions also include options for exporting and importing data for occupational coding. When the Export/Import for occupational coding menu items are chosen, data will be exported from/imported into multiple datasets. The resulting files will be a “pair” of macro-enabled Excel files for each questionnaire language code found in the database, one primary file and a second identical copy of the file to be used for double coding. When national teams complete the occupation coding and verify double coding agreement (through the internal check macro within the file) only the primary coded file is imported into the database.

The PISA Imports menu option contains specialized procedures designed to extract data from files delivered by the various electronic sources: the student delivery system (SDS), the online school and teacher questionnaires, the open-ended coding system (OECS), and ACER Maple sample management system. The DME application creates a log file for each imported data file to record the action for each data element encountered. All invalid data values are replaced with designated missing values and a record of that activity is added to an internal log table within the database.

It is the Data Manager’s responsibility to schedule and coordinate the various activities associated with the collection, entry and validation of the data in the database. They are typically allowed eight weeks after the last administration of the survey to gather and integrate the collected data into the database, including time for the human scoring of the cognitive items, and to perform all checks on the integrity and consistency of the data. For this last task the DME application provides the ability to perform various checks on the database. Two of them, the validation check and the Unique ID check, rarely yield actionable results as all methods of integrating data into the database undergo a validation check at the point of entry, and each dataset is designed so that duplicate ID’s can also be detected and prevented from entry into the database.

The Record Consistency check is a series of individual reports that are designed and scripted by Data Management to assist national teams with verifying:

1. Consistency between the absence codes in the sampling dataset and each of the other student datasets to determine if a student marked as absent has data in a related dataset or vice versa.
2. Consistency between the student demographics in the sampling dataset and the Student Background Questionnaire dataset.
3. Consistency between the cognitive response data files and their corresponding OECS datasets to ensure that all respondents received codes for the open-ended items.
4. Consistency between the questionnaire datasets and the cognitive datasets (i.e., whether a student took both sessions of the assessment).
5. Data entry inconsistencies of paper-based instruments.
6. Identification of missing response or coded occupational data.
7. Counts of certain aspects of the database, such as number of students by language of survey.

8. Consistency for the School and Teacher datasets related to participation, questionnaire data, and sampling information.
9. Identify the contents of specific inner tables, such as the “ImportValueErrors”, which captured all conversions of invalid data values into missing values.

These reports can be downloaded from the application to an Excel file. The NDM must review all of the cases identified in each report and, for all cases except the cases flagged in the “ImportValueErrors” check, the NDM should resolve the noted discrepancies or provide an explanation for why they could not be resolved. In addition to the Double-key entry report in the Record Consistency check, which checks for mis-matched IDs across datasets, there is also a separate Double-Key Data Entry check in the DME that is to be executed for all paper instruments, including the Parent Questionnaire. The **Double-Key Data Entry check** identifies inconsistent data values entered across corresponding data sets, such as, SBP1 and SBP2, the datasets containing the student questionnaires as entered by Key Entry Operator 1 and Key Entry Operator 2. For this check, the NDM must resolve all discrepancies before proceeding to the next step.

When the NDM is satisfied that all data that could be collected has been properly placed in the database and all discrepancies have been resolved or explained, the DME provides an export function that will create a read-only copy of the database where any variables that are designated for suppression (e.g., Personally Identifiable Information) are set to null values. This export database, along with the annotated consistency report document and, for CBA countries, a set of zip files containing all the electronic files that were imported into the database, are submitted to Data Management via a secure FTP site.

Pre-processing – National Database and Corresponding Files

When data were submitted to the Data Management contractor, a series of pre-processing steps were performed on the data to ensure completeness of the database and accuracy of the data.

Data submission from countries included any “unprocessed” files, or files that the DME software was not able to import. Data Management made great efforts to recover as much of this data as possible by repairing the files or finding and importing into the database a usable version from the PISA Uploads Server. To specifically handle the unique cases observed in PISA 2022, an additional file recovery tool was developed to expedite data recovery.

Running the DME software’s Record Consistency Checks outlined above was one of the first quality control checks on the data submission. In the field, National Centres were required to run these checks frequently for data quality and consistency. Although National Centres were required to execute these checks on their data, the Data Management contractor also executed these DME consistency checks in early data processing as a quick and efficient way to verify the quality of the data received.

All sampling data (variables and values) was verified against approved sampling data from the sampling international contractor, Westat, at the student-level and, if applicable, at the teacher-level as well.

These checks, in addition to other internal checks for coding, missing data, and student/teacher tracking data alignment with approved sampling forms, were executed upon receipt of the data. Reported inconsistencies returned from these checks were compiled and sent to the National Centre for more information and/or further corrections to the data. If necessary, National Centres resubmitted their data to the Data Management contractor for any missing or incorrect information and documented any changes made to the database in the consistency check report file. When countries redelivered data, Data Management refreshed the existing database with the newly-received data from the National Centre and continued with the same pre-processing steps again – executing another round of consistency checks to be sure all issues were resolved and/or documented. This initial step of processing (i.e., returning data inconsistencies to the National Centres and receiving a revised database) was an iterative process of data

review and validation. Once issues were resolved or documented, the data continued to the next phase of the internal process – loading the database into the cleaning and verification software.

Data Processing and Cleaning System

Loading the SQLite database into the Processing and Cleaning System

With all pre-processing checks complete, the country's database advanced to the next phase of the process – data cleaning and verification. To reach the high-quality requirements of PISA technical standards, the Data Management contractor created an efficient.NET application that uses SQL and SAS to merge and process datasets.

During the processing phase, one or two analysts independently loaded each national databases into the processing software, focusing on one country at a time, to complete all necessary phases of quality assurance. Once complete, SAS and SPSS datasets were delivered to the country, and other contractors for review and analyses.

The first step in this process was to load the pre-processed national database, an SQLite database, into the ETS Data Management cleaning and verification software. With the initial load of the database, specific quality assurance checks were applied to the data. These checks ensured:

- The project database delivered by the country used the most up-to-date template provided by the Data Management team which included all necessary patch files applied to the database. For PISA 2022, patch files were released by ETS Data Management and applied to the SQLite database by the National Data Manager to address issues in the codebook for proper data capture in the DME software. For example, a patch may be issued if an item was misclassified as having 4 response options instead of 5.
- The country database had the correct profile as dictated by the international options (e.g., Financial Literacy, *Une Heure* form, etc.) selected by the country.
- The number of cases in the data files by country/language agreed with the sampling information collected by Westat.
- All values for variables that used a value scheme were contained by that value scheme. For example, a variable may have the valid values of 1, 3 and 5; yet, this quality assurance check would capture if an invalid value, e.g. "4", was entered in the data.
- Valid values that may have been miskeyed as missing values were verified by the country. For example, valid values for a variable might range from "1" to "100" and data entry personnel may have mistakenly entered a value of "99", intending to issue a value of "999". This is common with paper-based instruments. Each suspicious data point was investigated and resolved by the country.
- Response data that appeared to have no logical connection to other response data (e.g. school/parent records possessing no relation to any student records) were validated to ensure correct IDs are captured.

Cognitive Assessment Data Processing

Integration

After the initial load of data and completion of early processing checks, the database entered the next phase of processing: Integration. During this integration phase, data which was structured within the country project database to assist in data collection was restructured to facilitate data cleaning. At the end

of this step, a single dataset was produced for each of the respondent types: student, school, and teacher (where applicable). Additionally, Parent questionnaire data was merged with their child/student data.

During data processing, the integration phase was critical because the Data Management contractor was able to analyse the data collected within the context of the sampling information supplied by the sampling contractor. Using this sampling information –captured in the Student Data File and Teacher Data File – extensive quality control checks were applied to the data in this phase. Over 100 quality assurance checks were performed on the database. As a result of these quality assurance checks, a data quality report was generated and delivered to countries to resolve outstanding issues and inconsistencies. This report was known as the Data Integrity ("DI") Report.

In this report, the Data Management contractor provided specific information to countries, including the name of the check and the description of the check as well as specific information, such as student IDs, for the cases that proved to be inconsistent or incorrect against the check. These checks included (but were not limited to):

- Cognitive test (FORMCODE) variable was blank or not valid.
- Student was missing key data needed for sampling and processing.
- Student was not within the allowable age for the assessment.
- Student was not represented in the Sampling Data (Student Data File).
- Students was marked absent yet had a response record.
- Student's grade was lower than allowed.
- Student's assessment path misaligned with the multi-stage, adaptive design.
- A teacher was marked as a "non-participant," yet response data existed for that teacher.
- The DI report was packaged along with a series of other quality control reports (i.e., harmonisation report and validation report, see "Background Questionnaire Assessment Data Processing") for national team review. When reviewing the report, National Centre teams were asked to review flagged inconsistencies from the report and correct data issues in the national database. National teams were instructed to complete the report review and revision of data within a specific timeframe for resubmission to the Data Management contractor. Additionally, national teams documented all data revisions in the DI report and returned the report to the Data Management contractor for review.
- After receiving the revised database and all documentation, the Data Management contractor repeated the pre-processing phase to ensure no new errors were reported and, if no issues or errors were found, the Data Management analyst re-executed the Integration step. As with the pre-processing consistency checks phase, the Integration step might have required several iterations and updates to country data if issues persisted and were not addressed by the National Centre. Frequently, one-on-one consultations were needed between the National Centre and the Data Management to resolve issues.

In addition to quality assurance reporting, a series of important data processing steps occurred during the Integration phase:

- Item Cluster Analysis: For the purposes of data processing, it is often convenient to be able to disaggregate a single variable into a collection of variables. To this end, a respondent's single booklet number was generated as a collection of Boolean variables which signalled the item clusters that the participant was exposed to by design. Similarly, the individual item responses for a participant were interpreted and coded into a single variable which represented the item clusters that the participant appears to have been presented. An analysis was performed to detect any inconsistencies between information in the student delivery system and information in the sampling design. Any discrepancies discovered were resolved by contacting the appropriate contractors.

- Raw Response Data Capture: In the case of paper-based administration, individual student selections (e.g., A, B, C, D) to multiple-choice items were captured accurately. This was not necessarily true in the case of computer-based administrations. While the student delivery system captures a student's response, it does not capture data in a format that could be used to conduct distractor analysis. The web-elements that are saved during a computer administration were therefore processed and interpreted into variables comparable to the paper-based administration.
- Timing: The student delivery system captured timing data for each screen viewed by the respondent. During the integration step, these timing variables are merged to the country database.
- Process Data: The student delivery system also produced log files where process data could be extracted for further analysis. Process data including the total response time, response time to first action, number of visits, number of short visits, and the number of actions were extracted by specialized tools and then verified by the Data Management contractor through a series of quality control checks. Such quality control checks identified inconsistencies or situations of unreasonableness (e.g., duplicated records, out-of-range values, system or operational issues, total unit duration is higher than item time). Once inconsistent results were either resolved or explained, the data were provided to psychometric teams for further analysis.
- SDS Post-processing: Necessary changes in the student delivery system were sometimes detected after the platform was already in use. For example, a test item that was scored by the delivery system may have had an error in the interpretation of a correct response, which was corrected in post-processing. These and other issues were resolved by the delivery system's developers and new scored response data was processed, issued, and merged by the Data Management contractor.
- Multi-Stage Adaptive Testing: For both the Reading and Mathematics CBA, counts and percentages were produced for each country. Such counts identified the breakdown of each stage by performance to confirm that the student delivery platform's routing worked as expected during the assessment.

Scoring

After initial integration of the data, the next phase of data management processing involved parallel processes that occur with assessment data:

- Scoring of test responses captured in paper booklets.
- Treatment of CBA human-coded items.
- Additional checks of cognitive items.

Scoring overview

The goal of the PISA assessment is to ensure comparability of the assessment results across countries. As a result, scoring of the responses to the test items was a critical component of the data management processing. While scores were generated for computer-based responses automatically, no such scoring variables existed for paper-based components. This step in the process was dedicated to creating these variables and inserting the relevant student responses. The Data Management contractor implemented rules from coding guides developed by the Test Development team. The coding guides were organised in sections, or clusters, that outlined the value, or score, for each response. The Data Management contractor was not only responsible for generating the syntax to implement the scoring rules but was also responsible for implementing a series of quality assurance checks on the data to determine any violations in scoring and/or any missing information.

When missing scores were present in variables where data was expected, the Data Management contractor consulted with the National Centre regarding these missing data. If National Centres were able

to resolve these issues (e.g., student response information was mistakenly mis-coded or not entered into the DME software), information was provided to the Data Management team through the submission of an updated, or revised, DME database and the necessary steps for pre-processing/processing were completed. If the reported data inconsistencies were resolved, the scoring process was deemed complete, and the data proceeded to the next phase of processing.

The scoring variables also served as a valuable data quality check. If any items appeared to function unexpectedly (i.e., too difficult, too easy, or unusually high missing rates), further investigation was carried out to determine if a booklet printing or translation error occurred or if systematic errors were introduced during the administration, data load, or data entry.

Once the Integration and Scoring steps were complete, the next phase of data cleaning involved the validation of the background questionnaire data, i.e., harmonisation of national adaptations and verification of questionnaire response data.

Background Questionnaire Assessment Data Processing

Harmonisation

Harmonisation, or harmonised variables

As mentioned earlier, although standardisation across countries was needed, countries had the opportunity to modify, or adapt, background questionnaire variable stems and response categories to reflect national specificities or contexts. These adaptations are referred to as “national adaptations.” While able to capture country contexts, these adapted variables needed to be mapped into the corresponding international variable for cross-country comparison.

More specifically, harmonisation or harmonising variables is a process of mapping the national response categories of a particular variable into the international response categories so they can be compared and analysed across countries. Not every nationally adapted variable required harmonisation, but for those that required harmonisation, the Data Management team assisted the Background Questionnaire contractor with creating the harmonisation mappings for each country using SAS code. This code was implemented into the cleaning system to handle these national variables during processing.

Additionally, harmonisation consisted of mapping adaptations for national variables where there was a structural change, e.g. question stem and/or variable response category options differ from the international version (this could be in the form of an addition or deletion of a response option and/or modification to the intent of the question stem or response option – as observed in variable SC013Q01TA where the country may alter the stem in creating a national adaptation and request information on the “type” of school in addition to whether the school is public or private). For example, more response categories may have been added or deleted (e.g., a variable may have five response options/choices to the question, but with the national adaptation the variable may have been modified to only have four response options/choices as only 4 make sense for the country’s purposes); or perhaps two questions were merged.

Overview of the workflow

To capture the appropriate adaptation and harmonisation, changes to variables by national teams were proposed during the translation and adaptation process. National adaptations for questionnaire variables were agreed upon by the Background Questionnaire contractor. These discussions regarding adaptations happened in the negotiation phase between the country and the contractor as well as the translation verification contractor – prior to data submission to ETS. All changes and adaptations to questionnaire variables were captured in the Questionnaire Adaptation Sheet (QAS).

It was the role of the Background Questionnaire contractor to use the country's QAS file to approve national adaptations as well as any corresponding harmonisation mapping. The Data Management contractor also assisted the Background Questionnaire contractor in developing the harmonisation code for use in the cleaning and verification software. Throughout this process, it was the responsibility of the Background Questionnaire contractor, with the assistance of the translation verification contractor, to ensure the QAS was complete and reflected the country's intent and interpretation.

Issues surrounding national adaptations and/or the harmonisation code produced by the cleaning software, often, involved consultation with the national team as well as the Background Questionnaire contractor. Both the Background Questionnaire contractor and the national team were responsible for reviewing the harmonisation report produced by the Data Management contractor during processing to verify national adaptations and corresponding mappings. Requested updates or changes were documented in the harmonisation report, the country QAS file, and the cleaning system harmonisation code. As a result of updates, a new harmonisation report was generated and delivered to the national team and the Background Questionnaire contractor for final review and approval.

Validation

After the Harmonisation step, the next phase in data cleaning and verification involved executing a series of validation checks on the data for contractor and country review.

Validation overview

In addition to nationally adapted variables, the Data Management contractor collaborated with the Background Questionnaire contractor to develop a series of validation checks that were performed on the data following harmonisation.

Validation checks are a set of consistency checks that provide National Centres with more detail concerning extreme and/or inconsistent values in their data. Issues detected by these checks were displayed in a validation report, which was shared with countries and contactors to observe these inconsistencies and potentially make improvements for the next cycle of PISA. Consistent with PISA 2018, national teams did not make changes to revise these extreme and/or inconsistent values in the report. Rather, national teams were instructed to leave the data as-is and make recommendations for addressing these issues in the data collection process during the next phase from Field Trial to Main Survey, or the next cycle of PISA.

Generally, validation checks captured inconsistent student, school, and teacher data. For example, these checks captured an inconsistency between the total number of years teaching, and the number of years teaching at a particular school (TC00701); or an inconsistency in student data related to the number of class periods per week in maths and the allowable total class periods per week (ST059Q02). Throughout the PISA cycle, these validation checks often served as valuable feedback to check on the data quality.

Treatment of inconsistent and extreme values in PISA 2022 main survey data

Following the approach implemented in PISA 2015 and PISA 2018 for extreme and/or inconsistent values within national data, the Data Management contractor, the Background Questionnaire contractor, and the OECD agreed on the implementation of specific range restriction rules applied during data cleaning that would manage extreme and/or inconsistent values. These values would be invalidated across all country databases.

Building on the range restriction rules developed in PISA 2015 and used in PISA 2018, the following principles were observed in the special handling of these inconsistent and/or extreme values:

- In most cases where there was an inconsistency, the question considered 'more difficult' was invalidated since this was more likely to have been answered inaccurately (for example, a question that involved memory recall or cognitive evaluation by the respondent; or, if an inconsistency existed between age and seniority, the proposed rule may invalidate seniority, but keep "age.").
- Apply stringent consistency and validity checks while computing derived variables. With this principle, the original values may be kept, while the values for the derived variable may have applied an "invalid" rule.

The specific range restriction rules for PISA 2022 are presented in Table 12.3.

Derived variables

Code in SAS to create derived variables was generated by the Background Questionnaire contractor for implementation into the cleaning system at this step in the process. The code to create derived variables included routines for calculating these variables, treating missing data appropriately, adding variable labels, etc. This code was based on the Main Survey (MS) Data Analysis Plan that outlined the derived variables that were calculated from PISA MS data.

As further explained in the MS Analysis Plan, for all questions in the MS questionnaires, regardless of whether they served as a basis for derived variables or not, the international database contains item-level data as obtained from the delivery platform. For any derived variables, whenever possible, these were specified consistent with previous cycles of PISA. In terms of this alignment, the first choice was alignment with PISA 2012, to enable comparison on math-related variables. The second choice was alignment with PISA 2018. This aimed to strike a balance and stability across recent and future cycles. A list of PISA 2022 Main Survey non-item response theory (IRT) derived variables ("simple indices") is available in Table 12.2.

As part of quality control, all derivations were verified by the Background Questionnaire contractor. Any updates or recoding made to the derived variable code were completed, documented, and redelivered to the Data Management contractor for use in the cleaning system. Data files were refreshed to implement any changes to the code or the variables.

Deliverables

After all data processing steps were complete and all updates to the data were made by national teams to resolve any issues or inconsistencies, the final phase of data processing included the creation of deliverable files for specific contractors (e.g., Westat for Sampling, or ETS for Data Analysis) as well as the National Centres. Each data file deliverable required a unique specification of variables along with their designated ordering within the file.

In addition to the generation of files for contractors and National Centre use, the 'deliverables' step in the cleaning process contained critical additions to the data – such as the addition of proxy scores, plausible values, background questionnaire scales, and sampling weights (student and teacher). The dynamic feature of the cleaning system allowed for the Data Management contractor to generate customized files for delivery at specific phases of the project lifecycle.

To produce these customized files for specific clients at specific phases of the project, each deliverable required a separate series of checks and reviews in order to ensure all data were handled appropriately and all values were populated as expected.

Preparing files for public use and analysis

To prepare for the public release of PISA 2022 main survey data, the Data Management contractor provided data files in SPSS and SAS to National Centres and the OECD Secretariat in batch deliveries at

various review points during the main survey cycle. With the initial data deliveries of the main survey, the data files included preliminary sampling weights and proxy proficiency scores for analysis. These data were later updated to include final sampling weights, plausible values, and questionnaire indices.

During each of these phases of delivery, National Centres reviewed these data files and provided ETS Data Management with any comments and/or revisions to the data.

The following data files were delivered:

- The Student combined data file contained all student responses to the test items (raw and scored), background questionnaire items, and optional questionnaire items such as Parent Questionnaire, Well-Being (WB) Questionnaire, Information and Computer Technology Literacy Familiarity (ICT) Questionnaire. These files included all raw variables, questionnaire indices, student weights, replicate weights, and plausible values.
- The School data file contained all response data collected with the School Questionnaires. These files included all raw responses, school-level base weights, questionnaire indices, and other derived variables.
- The Teacher data file contained response data from the Teacher Questionnaire. These files included all raw responses, questionnaire indices, derived variables, and teacher weights.
- The Financial literacy data file contained response data from the financial literacy cognitive and background questionnaire items. These files included all raw variables, questionnaire indices, sampling weights, replicate weights, and plausible values.
- The Masked international database, which combined the data from all participating countries. To preserve country anonymity in this file, key identifying variables were masked following specific guidelines from the OECD Secretariat that included issuing alternate codes or required special handling for country identifiers.
- The preliminary, national version of the Public Use File (PUF) was produced toward the end of the PISA 2022 main survey and provided the National Centre with the opportunity to review their data before the final public release. These data included all country-requested variable suppressions. More information about country-level variable suppressions is included in Table 12.3.

In addition to these data files, a series of analysis reports were produced by the Data Analysis team and delivered by the Data Management contractor to National Centres for quality control, data validation, and further national analyses. These reports were also used to evaluate the plausibility of the distributions of background characteristics and the performance results by subgroups, especially evaluating the extent to which they agree with expectations based on external or historical information. These reports included:

- BQ Crosstabs: A report containing frequencies of numerical, categorical variables from the country's Background Questionnaire (BQ). To aide countries in reviewing their BQ variables for potential translation or coding errors, flagging for outliers as compared across countries were included in this report.
- BQ MSIGS: A report containing summary statistics for all numerical variables from the country's Background Questionnaire.
- BQ SDTs: A set of reports containing summary data tables that provided descriptive statistics for every categorical background variable in the respective country's PISA data file. For each country, the summary data tables included both international and country-specific background variables.
- Codebook Descriptives Report: A report that includes 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 and includes descriptive statistics for all variables.
- Cognitive Summary Analysis Reports: A comprehensive report that included a series of key statistics and flags across item analysis (IA), coding reliability, and item response theory (IRT)

reports to identify items that, based on the empirical data, are most likely to require careful review and feedback by national teams.

- Item Analysis Reports: A set of reports that provided summary information about the response types given by the respondents to the cognitive items. They contained, for each country, various statistics (e.g., count, percent, mean cluster score) of students choosing each option for multiple-choice items or the percent of individuals receiving each score in the scoring guide for the constructed-response items.

The Public Use File - Included Records

When preparing for the final public use file (PUF), the following records were included in the database:

Student files

- Includes one records per respondent[1] that met the international target population definition and that passed validation, adjudication, and weighting.

School files

- Includes one record per participating school – specifically, one record for any school with a student included in the PISA sample regardless of whether the school returned the School Questionnaire.

Teacher files

- Include one record for each teacher that met the international target population definition and that passed validation, adjudication, and weighting[2].

Financial literacy student files

- One record per student respondent that met the international target population definition and that passed validation, adjudication, and weighting; and that responded to a cognitive form that included Financial Literacy items (Forms 67 – 74), or included Mathematics and reading items (Forms 1-12).

Categorising missing data

Within the data files, the coding of the data distinguishes between six different types of missing data:

1. System Missing/Blank – used to indicate that the respondent was not presented the question according to the survey design or ended the questionnaire early, or data loss.
2. No Response – 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.
3. Invalid – 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.
4. Not Applicable – used to indicate in the questionnaire that the question was not asked by design or could not be determined due to a printing problem or torn booklet, or due to within-construct matrix sampling design. In the cognitive data, it is used to indicate that the question was dropped/deleted during item calibration and not used during scaling.
5. Valid Skip – used in the questionnaire data to indicate that the question was not answered because a response to an earlier question directed the respondent to skip the question.

6. Not Reached – used in the cognitive scored variables to indicate that a student was unlikely to have seen the question and the response should be treated as such.

Data management and confidentiality, variable suppressions

During the PISA 2022 cycle, some country regulations and laws restricted the sharing of certain data with other countries. The key goal of such disclosure control is to prevent the accidental or intentional identification of individuals in the release of data. However, suppression of information or reduction of detail could impact the analytical utility of the data. Therefore, both goals must be carefully balanced. As a general directive for PISA 2022, the OECD requested that all countries make available the largest permissible set of information at the highest level of disaggregation possible.

Each country was required to provide early notification of any rules affecting the disclosure and sharing of PISA sampling, operational or response data. Furthermore, each country was responsible for implementing any additional confidentiality measures in the database before delivery to the Consortium. Most importantly, any confidentiality edits that changed the response values had to be applied prior to submitting data in order to work with identical values during processing, cleaning and analysis. The DME software only supported the suppression of entire variables. All other measures were implemented under the responsibility of the country via the export/import functionality or by editing individual data cells.

With the delivery of the data from the National Centre, the Data Management team reviewed a detailed document of information that included any implemented or required confidentiality practices to evaluate the impact on the data management cleaning and analysis processes. Country suppression requests generally involved specific variables that violate confidentiality and anonymity of student, school, and/or teacher data. To suppress data for the public use files, an invalid code was applied during the final step of data file creation in the cleaning system[3]. A listing of suppressions at the country variable-level is in Table 12.4.

Notes

1. To be considered a “respondent” the student must have responded to at least half of the number of test items in his or her booklet/form; or at least one test item response and a minimum number of responses to the student background questionnaire.

2. Teachers who were absent, excluded, or refused to participate in the session may be marked as a “non-participant.”

3. PISA national participants also had the opportunity to request a withdrawal of data. These requests were managed by the OECD and implemented by the Data Management contractor. The withdrawal of data involves removing data (e.g., records from specific regions) from data files and reports (including public-use files) for country-specific reasons. The request to withdrawal data required thorough discussion with the OECD and approval.

Chapter 12 tables

Table 12.1. PISA Variable Naming Convention

First Character	Second Character	Next Three Characters	Next Three Characters	Last Character
Indicates whether the variable is derived from the paper- or computer-based assessment	Indicates the cognitive domain for the related item	Is a unique numeric item identifier within each domain	Include of a "Q" and a two-digit numeric item part code.	Indicates additional information of the type of information captured.
P for **paper-based items** (Note: some of the paper-based reading and science trend items do not have "P" as the first character and, instead, may begin with "R" or "S" – see "Second Character" column) **C** for **computer-based items** (Note: Creative Thinking items do not have "C" as the first character, these variables begin with "T" – see "Second Character" column) **D** for **computer-based, human-coded items**	**M** for **Mathematics trend items** **MA** for **Mathematics new items** **R** for **Reading trend items** **S** for **Science trend items** **F** for **Financial Literacy items** **T** for **Creative Thinking items**			**S, SA, SB, SC, etc.** for the **scored response** C for a human-coded computer-based code R, RA, RB, RC, etc. for the actual response TT for the total timing F for the time to first action A for the number of actions V for the number of visits VS for the number of short visits

Table 12.2. PISA Non-IRT Derived Variables Code

Refer to <link> to view this table on line.

Table 12.3. PISA 2022 Range Restrict Code

Sequence	Dataset STU, SCH, TCH)	Description Code	SAS Code
STUDENT			
1	STU	INVALIDATE IF NUMBER FOR AN INDIVIDUAL'S WEIGHT IS NEGATIVE.	IF ((WB151Q01HA < 30) OR (WB151Q01HA > 250)) AND (NOT MISSING(WB151Q01HA)) THEN WB151Q01HA=.I;
2	STU	INVALIDATE IF NUMBER FOR AN INDIVIDUAL'S HEIGHT IS NEGATIVE.	IF ((WB152Q01HA < 90) OR (WB152Q01HA > 230)) AND (NOT MISSING(WB152Q01HA)) THEN WB152Q01HA=.I;
3	STU	INVALIDATE IF NUMBER FOR AN INDIVIDUAL'S CLOSE FRIENDS IS MORE THAN 50. (LISTED IN MAT'S EMAIL FROM 4/8/19 BUT NOT IN THIS EXCEL FILE)	IF (WB156Q01HA > 50) THEN WB156Q01HA =.I;
4	STU	INVALIDATE IF NUMBER OF CLASS PERIODS PER WEEK IN MATHEMATICS LESSONS (ST059Q01TA) IS NEGATIVE OR GREATER THAN 75	IF (ST059Q01TA > 75 OR ST059Q01TA < 0) AND NOT MISSING(ST059Q01TA) THEN ST059Q01TA =.I;
5	STU	INVALIDATE IF NUMBER OF TOTAL CLASS PERIODS IN A WEEK (ST059Q02JA) IS	IF (ST059Q02JA > 120 OR ST059Q02JA < 0) AND NOT MISSING(ST059Q02JA) THEN ST059Q02JA =.I;

		NEGATIVE OR GREATER THAN 120 OR LESS THAN 10.	
6	STU	INVALIDATE IF A CHILD'S ISCED LEVEL EQUALS 2 AND SELECTS THAT HE OR SHE HAS REPEATED ISCED 3 ONCE OR MULTIPLE TIMES	IF INT(ISCEDP/100)=2 AND (ST127Q03TA=2 OR ST127Q03TA=3) THEN ST127Q03TA =.I;
SCHOOL			
1	SCH	INVALIDATE IF TOTAL NUMBER OF COMPUTERS (SC004Q02TA) IS NEGATIVE.	IF (SC004Q02TA < 0) AND NOT MISSING(SC004Q02TA) THEN SC004Q02TA =.I;
2	SCH	INVALIDATE IF TOTAL NUMBER OF COMPUTERS (SC004Q03TA) IS NEGATIVE.	IF (SC004Q03TA < 0) AND NOT MISSING(SC004Q03TA) THEN SC004Q03TA =.I;
3	SCH	INVALIDATE IF TOTAL NUMBER OF WHITEBOARDS (SC004Q05NA) IS NEGATIVE.	IF (SC004Q05NA < 0) AND NOT MISSING(SC004Q05NA) THEN SC004Q05NA =.I;
4	SCH	INVALIDATE IF TOTAL NUMBER OF DATA PROJECTORS (SC004Q06NA) IS NEGATIVE.	IF (SC004Q06NA < 0) AND NOT MISSING(SC004Q06NA) THEN SC004Q06NA =.I;
5	SCH	INVALIDATE IF TOTAL NUMBER OF COMPUTERS (SC004Q07NA) IS NEGATIVE.	IF (SC004Q07NA < 0) AND NOT MISSING(SC004Q07NA) THEN SC004Q07NA =.I;
6	SCH	INVALIDATE IF TOTAL NUMBER OF TABLETS OR E-BOOK READERS (SC004Q08JA) IS NEGATIVE.	IF (SC004Q08JA < 0) AND NOT MISSING(SC004Q08JA) THEN SC004Q08JA =.I;
7	SCH	INVALIDATE IF NUMBER OF DESKTOP OR LAPTOP COMPUTERS CONNECTED TO THE INTERNET (SC004Q03TA) IS GREATER THAN THE NUMBER OF DESKTOP OF LAPTOP COMPUTERS AVAILABLE TO STUDENTS (SC004Q02TA).	IF SC004Q03TA > SC004Q02TA AND NOT MISSING(SC004Q02TA) THEN SC004Q03TA =.I;
8	SCH	INVALIDATE IF TOTAL NUMBER OF FULL-TIME TEACHERS (SC018Q01TA01) IS NEGATIVE.	IF (SC018Q01TA01 < 0) AND NOT MISSING(SC018Q01TA01) THEN SC018Q01TA01 =.I;
9	SCH	INVALIDATE IF NUMBER OF FULL-TIME CERTIFIED TEACHERS (SC018Q02TA01) IS NEGATIVE	IF (SC018Q01TA02 < 0) AND NOT MISSING(SC018Q01TA02) THEN SC018Q01TA02 =.I;
10	SCH	INVALIDATE IF NUMBER OF FULL-TIME CERTIFIED TEACHERS (SC018Q02TA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC018Q01TA01).	IF SC018Q02TA01 > SC018Q01TA01 AND NOT MISSING(SC018Q01TA01) THEN SC018Q02TA01 =.I;
11	SCH	INVALIDATE IF NUMBER OF FULL-TIME BACHELOR DEGREE TEACHERS (SC018Q08JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC018Q01TA01).	IF SC018Q08JA01 > SC018Q01TA01 AND NOT MISSING(SC018Q01TA01) THEN SC018Q08JA01 =.I;
12	SCH	INVALIDATE IF NUMBER OF FULL-TIME MASTER'S DEGREE TEACHERS (SC018Q09JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC018Q01TA01).	IF SC018Q09JA01 > SC018Q01TA01 AND NOT MISSING(SC018Q01TA01) THEN SC018Q09JA01 =.I;
13	SCH	INVALIDATE IF NUMBER OF FULL-TIME DOCTORAL DEGREE TEACHERS (SC018Q10JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC018Q01TA01).	IF SC018Q10JA01 > SC018Q01TA01 AND NOT MISSING(SC018Q01TA01) THEN SC018Q10JA01 =.I;
14	SCH	INVALIDATE IF NUMBER OF PART TIME CERTIFIED TEACHERS (SC018Q02TA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC018Q01TA02).	IF SC018Q02TA02 > SC018Q01TA02 AND NOT MISSING(SC018Q01TA02) THEN SC018Q02TA02 =.I;
15	SCH	INVALIDATE IF NUMBER OF PART TIME BACHELOR DEGREE TEACHERS (SC018Q08JA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC018Q01TA02).	IF SC018Q08JA02 > SC018Q01TA02 AND NOT MISSING(SC018Q01TA02) THEN SC018Q08JA02 =.I;

16	SCH	INVALIDATE IF NUMBER OF PART TIME MASTER'S DEGREE TEACHERS (SC018Q09JA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC018Q01TA02).	IF SC018Q09JA02 > SC018Q01TA02 AND NOT MISSING(SC018Q01TA02) THEN SC018Q09JA02 =.I;
17	SCH	INVALIDATE IF NUMBER OF PART TIME DOCTORAL DEGREE TEACHERS (SC018Q10JA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC018Q01TA02).	IF SC018Q10JA02 > SC018Q01TA02 AND NOT MISSING(SC018Q01TA02) THEN SC018Q10JA02 =.I;
18	SCH	INVALIDATE IF TOTAL NUMBER OF FULL-TIME MATHEMATICS TEACHERS (SC182Q01WA01) IS NEGATIVE.	IF (SC182Q01WA01 < 0) AND NOT MISSING(SC182Q01WA01) THEN SC182Q01WA01 =.I;
19	SCH	INVALIDATE IF NUMBER OF FULL-TIME CERTIFIED MATHEMATICS TEACHERS (SC182Q06WA01) EXCEEDS TOTAL NUMBER OF FULL-TIME MATHEMATICS TEACHERS (SC182Q01WA01).	IF SC182Q06WA01 > SC182Q01WA01 AND NOT MISSING(SC182Q01WA01) THEN SC182Q06WA01 =.I;
20	SCH	INVALIDATE IF NUMBER OF FULL-TIME MATHEMATICS BACHELOR DEGREE TEACHERS (SC182Q07JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA01).	IF SC182Q07JA01 > SC182Q01WA01 AND NOT MISSING(SC182Q01WA01) THEN SC182Q07JA01 =.I;
21	SCH	INVALIDATE IF NUMBER OF FULL-TIME MATHEMATICS TEACHERS WITH BACHELOR DEGREE AND MATH MAJOR (SC182Q08JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA01).	IF SC182Q08JA01 > SC182Q01WA01 AND NOT MISSING(SC182Q01WA01) THEN SC182Q08JA01 =.I;
22	SCH	INVALIDATE IF NUMBER OF FULL-TIME MATHEMATICS TEACHERS WITH BACHELOR DEGREE AND PEDGOGY QUALIFCATION (SC182Q09JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA01).	IF SC182Q09JA01 > SC182Q01WA01 AND NOT MISSING(SC182Q01WA01) THEN SC182Q09JA01 =.I;
23	SCH	INVALIDATE IF NUMBER OF FULL-TIME MATHEMATICS ISCED 5 TEACHERS (SC182Q10JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA01).	IF SC182Q10JA01 > SC182Q01WA01 AND NOT MISSING(SC182Q01WA01) THEN SC182Q10JA01 =.I;
24	SCH	INVALIDATE IF TOTAL NUMBER OF PART TIME MATHEMATICS TEACHERS (SC182Q01WA02) IS NEGATIVE.	IF (SC182Q01WA02 < 0) AND NOT MISSING(SC182Q01WA02) THEN SC182Q01WA02 =.I;
25	SCH	INVALIDATE IF NUMBER OF PART TIME CERTIFIED TEACHERS (SC182Q06WA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC182Q01WA02).	IF SC182Q06WA02 > SC182Q01WA02 AND NOT MISSING(SC182Q01WA02) THEN SC182Q06WA02 =.I;
26	SCH	INVALIDATE IF NUMBER OF PART TIME MATHEMATICS BACHELOR DEGREE TEACHERS (SC182Q07JA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC182Q01WA02).	IF SC182Q07JA02 > SC182Q01WA02 AND NOT MISSING(SC182Q01WA02) THEN SC182Q07JA02 =.I;
27	SCH	INVALIDATE IF NUMBER OF PART TIME MATHEMATICS TEACHERS WITH BACHELOR DEGREE AND MATH MAJOR (SC182Q08JA02) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA02).	IF SC182Q08JA02 > SC182Q01WA02 AND NOT MISSING(SC182Q01WA02) THEN SC182Q08JA02 =.I;
28	SCH	INVALIDATE IF NUMBER OF PART TIME MATHEMATICS TEACHERS WITH BACHELOR DEGREE AND PEDAGOGY QUALIFICATION (SC182Q09JA02) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA02).	IF SC182Q09JA02 > SC182Q01WA02 AND NOT MISSING(SC182Q01WA02) THEN SC182Q09JA02 =.I;

29	SCH	INVALIDATE IF NUMBER OF PART TIME MATHEMATICS ISCED 5 TEACHERS (SC182Q10JA02) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA02).	IF SC182Q10JA02 > SC182Q01WA02 AND NOT MISSING(SC182Q01WA02) THEN SC182Q10JA02 =.I;
30	SCH	INVALIDATE IF SUM OF FUNDING PERCENTAGES IS LESS THAN 98% OR GREATER THAN 102% (SC016Q01TA + SC016Q02TA + SC016Q03TA + SC016Q04TA).	IF SUM(SC016Q01TA,SC016Q02TA,SC016Q03TA,SC016Q04TA) > 102 OR SUM(SC016Q01TA,SC016Q02TA,SC016Q03TA,SC016Q04TA) < 98 THEN DO; SC016Q01TA =.I;SC016Q02TA =.I;SC016Q03TA =.I;SC016Q04TA =.I;
31	SCH	INVALIDATE IF PERCENTAGE OF TEACHING STAFF (SC025Q01NA) IS GREATER THAN 100%.	IF SC025Q01NA>100 THEN SC025Q01NA =.I;
32	SCH	INVALIDATE IF PERCENTAGE OF MATHEMATICS TEACHER STAFF (SC025Q02NA) IS GREATER THAN 100%.	IF SC025Q02NA>100 THEN SC025Q02NA =.I;
33	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WITH <HERITAGE LANGUAGE> DIFFERENT THAN <TEST LANGUAGE> (SC211Q01JA) IS GREATER THAN 100%.	IF SC211Q01JA>100 THEN SC211Q01JA =.I;
34	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WITH SPECIAL LEARNING NEEDS (SC211Q02JA) IS GREATER THAN 100%.	IF SC211Q02JA>100 THEN SC211Q02JA =.I;
35	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS FROM DISADVANTAGED HOMES (SC211Q03JA) IS GREATER THAN 100%.	IF SC211Q03JA>100 THEN SC211Q03JA =.I;
36	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WHO ARE IMMIGRANTS (SC211Q04JA) IS GREATER THAN 100%.	IF SC211Q04JA>100 THEN SC211Q04JA =.I;
37	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WHOSE PARENTS ARE IMMIGRANTS (SC211Q05JA) IS GREATER THAN 100%.	IF SC211Q05JA>100 THEN SC211Q05JA =.I;
38	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WHO ARE REFUGEES (SC211Q06JA) IS GREATER THAN 100%.	IF SC211Q06JA>100 THEN SC211Q06JA =.I;
39	SCH	INVALIDATE IF PERCENTAGE OF PARENTS THAT INITIATED DISCUSSION ON CHILD'S PROGRESS (SC064Q01TA) IS GREATER THAN 100%.	IF SC064Q01TA>100 THEN SC064Q01TA =.I;
40	SCH	INVALIDATE IF PERCENTAGE OF PARENTS WHERE TEACHER-INITIATED DISCUSSION ON CHILD'S PROGRESS (SC064Q02TA) IS GREATER THAN 100%.	IF SC064Q02TA>100 THEN SC064Q02TA =.I;
41	SCH	INVALIDATE IF PERCENTAGE OF PARENTS PARTICIPATED IN SCHOOL GOVERNMENT (SC064Q03TA) IS GREATER THAN 100%.	IF SC064Q03TA>100 THEN SC064Q03TA =.I;
42	SCH	INVALIDATE IF PERCENTAGE OF PARENTS THAT VOLUNTEERED IN EXTRACURRICULAR ACTIVITIES (SC064Q04NA) IS GREATER THAN 100%.	IF SC064Q04NA>100 THEN SC064Q04NA =.I;
43	SCH	INVALIDATE IF PERCENTAGE OF PARENTS THAT INITIATED DISCUSSION ON CHILD'S BEHAVIOR (SC064Q05WA) IS GREATER THAN 100%.	IF SC064Q05WA>100 THEN SC064Q05WA =.I;
44	SCH	INVALIDATE IF PERCENTAGE OF PARENTS WHERE TEACHER-INITIATED DISCUSSION ON CHILD'S BEHAVIOR (SC064Q06WA) IS GREATER THAN 100%.	IF SC064Q06WA>100 THEN SC064Q06WA =.I;

45	SCH	INVALIDATE IF PERCENTAGE OF PARENTS THAT ASSISTED IN FUNDRAISING (SC064Q07WA) IS GREATER THAN 100%.	IF SC064Q07WA>100 THEN SC064Q07WA =.I;
46	SCH	INVALIDATE IF TOTAL NUMBER OF BOYS (SC002Q01TA) AND TOTAL NUMBER OF GIRLS (SC002Q02TA) ARE BOTH ZERO.	IF SC002Q01TA=0 AND SC002Q02TA=0 THEN DO; SC002Q01TA =.I; SC002Q02TA=.I; END;
47	SCH	INVALIDATE IF TOTAL NUMBER OF STUDENTS IN MODAL GRADE (SC004Q01TA) IS GREATER THAN TOTAL NUMBER OF STUDENTS (SC002Q01TA + SC002Q02TA).	IF SC004Q01TA > SUM(SC002Q01TA,SC002Q02TA) THEN SC004Q01TA =.I;
48	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WITH MARKS AT OR ABOVE (SC178Q01JA) AND BELOW PASSING (SC178Q02JA) IS GREATER THAN 100%.	IF SUM(SC178Q01JA + SC178Q02JA) >100 THEN DO; SC025Q01NA =.I; SC178Q01JA=.I; SC178Q02JA=.I; END;
49	SCH	INVALIDATE IF TOTAL NUMBER OF NON-TEACHING STAFF (SC168Q01JA) IS NEGATIVE.	IF (SC168Q01JA < 0) AND NOT MISSING(SC168Q01JA) THEN SC168Q01JA =.I;
50	SCH	INVALIDATE IF TOTAL NUMBER OF NON-TEACHING STAFF (SC168Q02JA) IS NEGATIVE.	IF (SC168Q02JA < 0) AND NOT MISSING(SC168Q02JA) THEN SC168Q02JA =.I;
51	SCH	INVALIDATE IF TOTAL NUMBER OF NON-TEACHING STAFF (SC168Q03TA) IS NEGATIVE.	IF (SC168Q03JA < 0) AND NOT MISSING(SC168Q03JA) THEN SC168Q03JA =.I;
52	SCH	INVALIDATE IF TOTAL NUMBER OF NON-TEACHING STAFF (SC168Q04TA) IS NEGATIVE.	IF (SC168Q04JA < 0) AND NOT MISSING(SC168Q04JA) THEN SC168Q04JA =.I;
53	SCH	INVALIDATE IF TOTAL NUMBER OF FOREIGN LANGUAGES (SC174Q01JA) IS NEGATIVE.	IF (SC174Q01JA < 0) AND NOT MISSING(SC174Q01JA) THEN SC174Q01JA =.I;
54	SCH	INVALIDATE IF TOTAL NUMBER OF DAYS (SC213Q01JA) IS NEGATIVE.	IF (SC213Q01JA < 0) AND NOT MISSING(SC213Q01JA) THEN SC213Q01JA =.I;
55	SCH	INVALIDATE IF TOTAL NUMBER OF DAYS (SC213Q02JA) IS NEGATIVE.	IF (SC213Q02JA < 0) AND NOT MISSING(SC213Q02JA) THEN SC213Q02JA =.I;
56	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WITH MARKS AT OR ABOVE (SC178Q01JA) AND BELOW PASSING (SC178Q02JA) ARE BOTH ZERO.	IF (SC178Q01JA = 0 AND SC178Q02JA = 0) THEN DO; SC178Q01JA=.I; SC178Q02JA=.I; END;
57	SCH	INVALIDATE IF TOTAL NUMBER OF DAYS (SC213Q01JA) IS NEGATIVE OR >1000.	IF (SC213Q01JA < 0) AND NOT MISSING(SC213Q01JA) THEN SC213Q01JA =.I; IF (SC213Q01JA >1000 THEN SC213Q01JA =.I;
58	SCH	INVALIDATE IF TOTAL NUMBER OF DAYS (SC213Q02JA) IS NEGATIVE OR >1000.	IF (SC213Q02JA < 0) AND NOT MISSING(SC213Q02JA) THEN SC213Q02JA =.I; IF (SC213Q02JA >1000 THEN SC213Q02JA =.I;

59	SCH	(SC175Q01JA, SC175Q02JA) THE MINUTES PER CLASS PERIOD SHOULD SET TO 1-120	IF (SC175Q01JA < 1) AND NOT MISSING(SC175Q01JA) THEN SC175Q01JA =.I;
			IF (SC175Q01JA >120 THEN SC175Q01JA =.I;
			IF (SC175Q02JA < 1) AND NOT MISSING(SC175Q02JA) THEN SC175Q02JA =.I;
			IF (SC175Q02JA >120 THEN SC175Q02JA =.I;

TEACHER			
1	TCH	INVALIDATE IF NUMBER OF YEARS TEACHING AT SCHOOL (TC007Q01NA) EXCEEDS REPORTED AGE (TC002Q01NA) MINUS 15.	IF TC007Q01NA > (TC002Q01NA - 15) AND NOT MISSING(TC002Q01NA) THEN TC007Q01NA =.I;
2	TCH	INVALIDATE IF TOTAL NUMBER OF YEARS TEACHING (TC007Q02NA) EXCEEDS REPORTED AGE (TC002Q01NA) MINUS 15.	IF TC007Q02NA > (TC002Q01NA - 15) AND NOT MISSING(TC002Q01NA) THEN TC007Q02NA =.I;
3	TCH	INVALIDATE IF YEARS WORKING AS A TEACHER IN TOTAL (TC007Q02NA) IS	IF TC007Q01NA > TC007Q02NA AND NOT MISSING(TC007Q02NA) THEN TC007Q01NA =.I;

		LESS THAN YEARS WORKING AS A TEACHER IN THIS SCHOOL (TC007Q01NA).	
4	TCH	INVALIDATE IF SUM OF TEACHER EDUCATION OR TRAINING PROGRAMME OR OTHER PROFESSIONAL QUALIFICATION IS LESS THAN 98% OR GREATER THAN 102% (TC203Q01HA + TC203Q02HA +TC203Q03HA)	IF SUM(TC203Q01HA, TC203Q02HA, TC203Q03HA) > 102 OR SUM(TC203Q01HA, TC203Q02HA, TC203Q03HA) < 98 THEN DO; TC203Q01HA =.I; TC203Q02HA=.I; TC203Q03HA =.I;
5	TCH	INVALIDATE IF SUM OF TEACHER EDUCATION OR TRAINING PROGRAMME OR OTHER PROFESSIONAL QUALIFICATION DURING THE LAST 12 MONTHS IS LESS THAN 98% OR GREATER THAN 102% (TC204Q01HA + TC204Q02HA +TC204Q03HA)	IF SUM(TC204Q01HA, TC204Q02HA, TC204Q03HA) > 102 OR SUM(TC203Q01HA, TC203Q02HA, TC203Q03HA) < 98 THEN DO; TC204Q01HA =.I; TC204Q02HA=.I; TC204Q03HA =.I;
6	TCH	INVALIDATE IF NUMBER OF DAYS (TC257Q01JA) IS NEGATIVE.	IF (TC257Q01JA < 0) AND NOT MISSING(TC257Q01JA) THEN TC257Q01JA =.I;

Table 12.4. PISA Country Variable Suppressions

Country Variable Suppression	
Austria	
GRADE	SC016Q01TA
OCOD1 (2-digit)	SC016Q02TA
OCOD2 (2-digit)	SC016Q03TA
PROGN	SC016Q04TA
SC001Q01TA (recoding)	SCHLTYPE
SC002Q01TA	SCHSIZE (recoding)
SC002Q02TA	ST001D01T (recoding)
SC004Q01TA	STRATUM
SC014Q01TA	
Belgium (French/German)	
ST003D02T	
Canada	
CLSIZE	SC176Q01JA
MCLSIZE	SC182Q01WA01
SC002Q01TA	SC182Q01WA02
SC002Q02TA	SC182Q06WA01
SC003Q01TA	SC182Q06WA02
SC004Q01TA	SC182Q07JA01
SC018Q01TA01	SC182Q07JA02
SC018Q01TA02	SC182Q08JA01
SC018Q02TA01	SC182Q08JA02
SC018Q02TA02	SC182Q09JA01
SC018Q08JA01	SC182Q09JA02
SC018Q08JA02	SC182Q10JA01
SC018Q09JA01	SC182Q10JA02
SC018Q09JA02	SCHSIZE
SC018Q10JA01	SMRATIO
SC018Q10JA02	STRATIO
SC168Q01JA	STRATUM
SC168Q02JA	TOTAT
SC168Q03JA	TOTMATH
SC168Q04JA	TOTSTAFF
Cyprus	
LANGTEST_COG	
LANGTEST_QQQ	
LANGTEST_QQQ	
SC001Q01TA	

Country Variable Suppression	
STRATUM	
Germany	
STRATUM	
Iceland	
GRADE	ST003D02T
SC002Q01TA	ST019AQ01T
SC002Q02TA	ST019BQ01T
SC004Q01TA	ST019CQ01T
SC013Q01TA	ST022Q01TA
SC014Q01TA	ST230Q01JA
ST001D01T	TOTAT
Israel	
STRATUM	
Italy	
REGION	
STRATUM	
Japan	
IMMIG	
Jordan	
STRATUM	
Macao	
LANGTEST_COG	
LANGTEST_PAQ	
LANGTEST_QQQ	
PRIVATESCH	
PROGN	
SC013Q01TA	
SCHLTYPE	
Malaysia	
SC012Q03TA	ST261Q03JA
SC012Q05TA	ST261Q09JA
SC012Q08JA	ST265Q03JA
SC012Q10JA	ST265Q04JA
SC012Q12JA	ST266Q02JA
ST038Q09JA	ST266Q03JA
ST038Q10JA	ST266Q04JA
ST038Q11JA	ST266Q05JA
ST261Q02JA	
Montenegro	
SC013Q01TA	
ST003D02T	
New Zealand	
SC002Q01TA	SC182Q06WA01
SC002Q02TA	SC182Q06WA02
SC004Q01TA	SC182Q07JA01
SC004Q02TA	SC182Q07JA02
SC018Q01TA01	SC182Q08JA01
SC018Q01TA02	SC182Q08JA02
SC018Q02TA01	SC182Q09JA01
SC018Q02TA02	SC182Q09JA02
SC018Q08JA01	SC182Q10JA01
SC018Q08JA02	SC182Q10JA02
SC018Q09JA01	SCHSIZE
SC018Q09JA02	TOTAT
SC018Q10JA01	TOTMATH
SC018Q10JA02	TOTSTAFF
SC182Q01WA01	WB151Q01HA

Country Variable Suppression	
SC182Q01WA02	WB152Q01HA
Norway	
CLSIZE	SC018Q08JA01
GRADE	SC018Q08JA02
LANGTEST_COG	SC018Q09JA01
LANGTEST_QQQ	SC018Q09JA02
MCLSIZE	SC018Q10JA01
PRIVATESCH	SC018Q10JA02
PROADMIN	SC168Q01JA
PROATCE	SC168Q02JA
PROMGMT	SC168Q03JA
PROOSTAF	SC168Q04JA
PROPAT6	SC182Q01WA01
PROPAT7	SC182Q01WA02
PROPAT8	SC182Q06WA01
PROPMATH	SC182Q06WA02
PROPSUPP	SC182Q07JA01
RATCMP1	SC182Q07JA02
RATCMP2	SC182Q08JA01
RATTAB	SC182Q08JA02
SC002Q01TA	SC182Q09JA01
SC002Q02TA	SC182Q09JA02
SC004Q01TA	SC182Q10JA01
SC004Q02TA	SC182Q10JA02
SC004Q03TA	SCHLTYPE
SC004Q05NA	SCHSIZE
SC004Q06NA	SMRATIO
SC004Q07NA	ST001D01T
SC004Q08JA	ST003D02T
SC012Q03TA	ST003D03T
SC013Q01TA	STRATIO
SC014Q01TA	TOTAT
SC016Q01TA	TOTMATH
SC016Q02TA	TOTSTAFF
SC016Q03TA	
SC016Q04TA	
SC018Q01TA01	
SC018Q01TA02	
SC018Q02TA01	
SC018Q02TA02	
Singapore	
LANGN	
OCOD1 (2-digit)	
OCOD2 (2-digit)	
SC211Q02JA	
SC211Q03JA	
Sweden	
GRADE	SC182Q06WA02
SC001Q01TA	SC182Q07JA01
SC002Q01TA	SC182Q07JA02
SC002Q02TA	SC182Q08JA01
SC004Q01TA	SC182Q08JA02
SC004Q02TA	SC182Q09JA01
SC004Q03TA	SC182Q09JA02
SC004Q08JA	SC182Q10JA01
SC013Q01TA	SC182Q10JA02
SC014Q01TA	SC211Q01JA
SC018Q01TA01	SC211Q02JA

Country Variable Suppression	
SC018Q01TA02	SC211Q03JA
SC018Q02TA01	SC211Q04JA
SC018Q02TA02	SC211Q05JA
SC018Q08JA01	SC211Q06JA
SC018Q08JA02	ST001D01T
SC018Q09JA01	ST003D02T
SC018Q09JA02	ST003D03T
SC018Q10JA01	ST021Q01TA
SC018Q10JA02	ST022Q01TA
SC182Q01WA01	ST126Q01TA
SC182Q01WA02	ST226Q01JA
SC182Q06WA01	
Thailand	
STRATUM	

Note: 1. Cyprus data are suppressed from the public use files. Information on data for Cyprus: https://oe.cd/cyprus-disclaimer.