10

Data Management Procedures

INTRODUCTION

In PISA, as in any international survey, a set of standard, data collection requirements guides the creation of an international database that allows for valid within-and-cross-country comparisons and inferences to be made. For both paper-based (PBA) and computer-based (CBA) assessments, these standard requirements are developed with three major goals in mind: consistency, precision and generalisability. In order to support these goals, data collection and management procedures are applied in a common and consistent way across all participants' data to ensure data quality. Even the smallest errors in data capture, coding, and/or processing may be difficult, if not impossible, to correct; thus, there is a critical need to 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 was conducted with the national representatives of participating countries in order to reflect country expectations in agreement with PISA 2015 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 on 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. These deviations may compromise data quality and/or render data corrupt, or unusable. For example, certain deviations from the standard testing procedures are particularly likely to 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 – the PISA timeline and the diversity of contexts in the administration of the assessment – 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 had to be generalised to suit the particular cognitive test instruments and background questionnaire instruments used in each participating country. As a result, a suite of products are provided to countries that include a comprehensive data management manual, training sessions, as well as a range of other materials, and in particular, the data management software designed to help National Project Managers (NPMs) and National Data Managers (NDMs) carry out in a consistent way data management tasks, prevent introduction of errors, and reduce the amount of effort and time in identifying and resolving data errors.

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 to ensure data quality in PISA 2015:

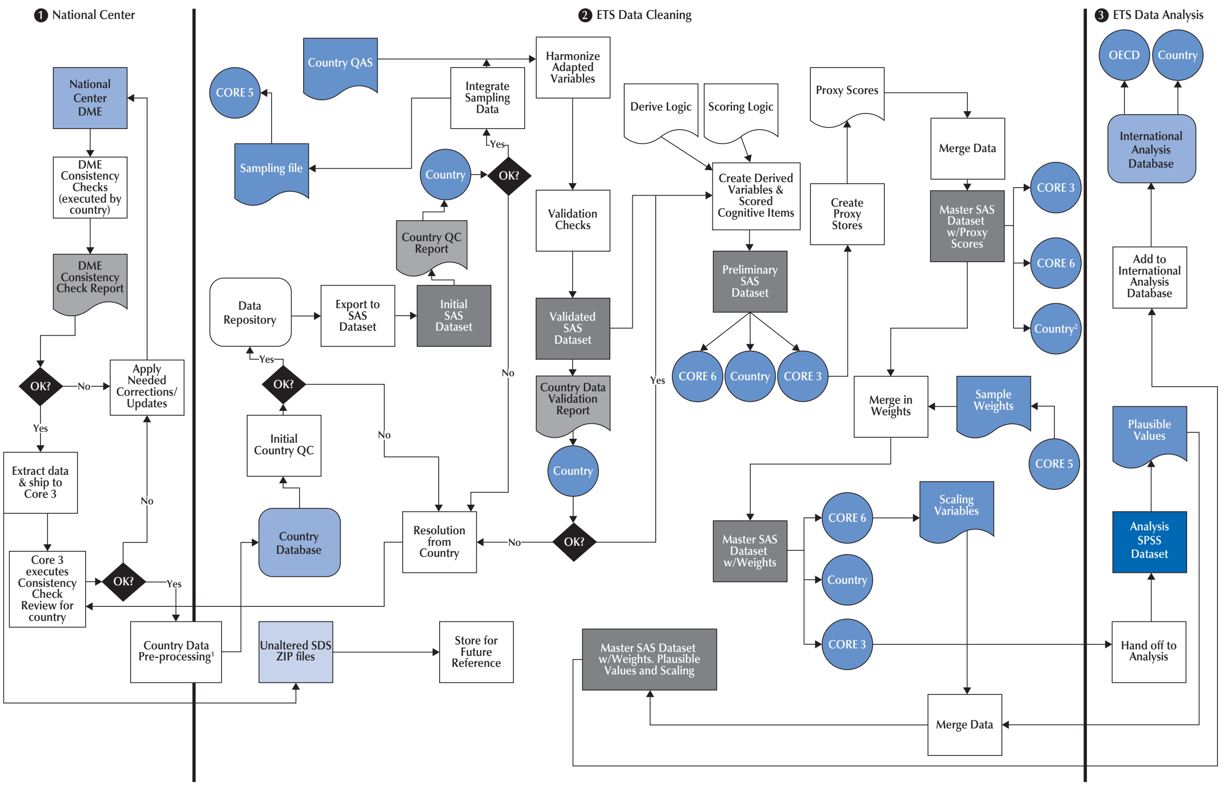
- standards, guidelines, and recommendations for data management within countries
- data management software, manuals and codebooks for National Centres
- hands-on data management training and support for countries during the national database building
- management, processing, and cleaning for data quality and verification at the international and national level
- preparation of analysis and dissemination of databases and reports for use by the contractors, OECD and the National Centres
- preparation of data products (e.g. Data Explorer, IDB Analyser) for dissemination to contractors, National Centres, the OECD, and the scientific community.

ETS Data Management and Analysis had overall responsibility for data management and relied on the following for information and consultation:

- ETS Project Management (Core 2 and Core 7): ETS Project Management provided contractors with overview information on country specifics including national options, timelines and testing dates, and support with country correspondence and deliverables planning.
- DIPF (Core 6): As the Background Questionnaire (BQ) experts, DIPF provided BQ scaling and indices, BQ data, support for questionnaire workflows and negotiations with National Centres concerning questionnaire national adaptations, harmonisation review, and BQ derived variables.
- Westat (Sampling) (Core 5): Leading the Sampling tasks for PISA, Westat provided review and quality control support with respect to sampling and weighting. Westat is instrumental in providing guidance for quality assurance checks with regard to national samples.
- Westat (Survey Operations) (Core 4): Key to the implementation of the PISA assessment in countries, Westat's Survey Operations team supported countries through the PISA 2015 cycle. In addition to organising PISA meetings, Westat was responsible for specific quality assurance of the implementation of the assessment and submission of data to the National Centres.
- OECD: The OECD provided support and guidance to all contractors with respect to the specific area of expertise. The OECD's review of data files and preliminary data products provided the ETS Data Management and Analysis teams with valuable information in the structure of the final deliverables.

Figure 10.1

Overview of the data management process



1. Additional chacks on data; data recovery processing; rescoring of cognitive items; timing data extraction; coding rehability export for analysis.
2. Interim databases delivery to country included proxy scopes.

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, or 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[1] tasks at the national level. Further, 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 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; manual capture of data after scoring; export/import of data required for coding (e.g. occupational coding); and data verification and validation with a series of consistency and validity checks. In order 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, or "DME." For PISA 2015, Figure 10.1 provides the workflow 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 2015 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 test administrator responses on a computer platform into electronic data files provided an opportunity for the accurate transcription of those responses and the collection of process data, including response actions and timing. It also presented a challenge to develop a system that accepted and processed these files and their variety of formats as well as supporting the manual entry of data from paper forms and booklets. To that end, the Data Management team acquired a license for the adaptation, use, and support of the Data Management Expert (DME) software, which had previously proved successful in the collection and management of the data for the survey for adult skills (PIAAC) under a separate contract.

The DME software 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 Windows, and does not require an internet connection to operate. It operates on a separate database file that has been constructed according to strict structural and relational specifications that define the data codebook. This codebook is a complete catalogue of all of the data variables to be collected and managed and the arrangement of these variables into well-defined datasets that correspond to the various instruments involved in the administration of the assessment. The DME software validates the structure of the codebook part of the database file and, if successful, creates the data tables within the same file for the collection and management of the response and derivative data.

With this process, the Data Management contractor first developed and tested a template of the international data codebook representing all the data to be collected across CBA and PBA countries without national adaptations. The datasets in this codebook also included those for all international options (such as financial literacy, teacher questionnaires, etc.) regardless of each country's mode or selected options. The national templates for each of the CBA countries are built upon this international template, using the questionnaire adaptations coded in the Questionnaire Adaptation Tool (QAT) and removing the datasets for PBA countries and the international options not implemented in the country. The national templates for each of the PBA countries consist of the international template with the CBA-specific datasets removed. The National Data Manager (NDM) for each PBA country is trained on and is responsible for implementing and testing the national adaptations to the delivered codebook.

The DME software provided three modes of entering data into the project database: imports of standard format files, imports of PISA specific archive files, and direct manual entry from paper forms and booklets. The standard format files are either Excel workbooks or CSV files and include such data as the results of the occupational coding. The PISA-specific files include the archive files that are generated by the student delivery system (SDS) software at the student level and the school and teacher questionnaire data files that are downloaded from the questionnaire web site by each NDM. The identification and extraction of data from these sources requires special programming within the DME software and supporting tables within the codebook files.

PBA countries performed direct manual entry into the system from paper forms and booklets. PBA data managers were required to program the codebook with the appropriate variables based on the booklet number and according to data management guidelines. Data entry was also required for the Parent Questionnaire when that option was selected by both PBA and CBA countries. An important feature of the DME software is the ability to create multiple copies of the project codebook for use on remote computers and to merge the databases created on each remote site into the master project database. This permits the establishment of a manageable processing environment based on a common codebook structure to guarantee the accurate and consistent transcription of the data.

The DME software can also produce a series of reports at any point during data collection, including: detection of records with the same identification information, validation of all data values against the codebook specifications, and a set of consistency checks defined and coded by the Data Management contractor. These checks provided information on the completeness of the data across datasets, identified inconsistent responses within each questionnaire, and reported on the overall status of the data collection process. At the conclusion of data collection and processing in each country, the NDM was required to either resolve or explain the discrepancies uncovered by these reports and submit the annotated reports along with the final database to the Data Management contractor.

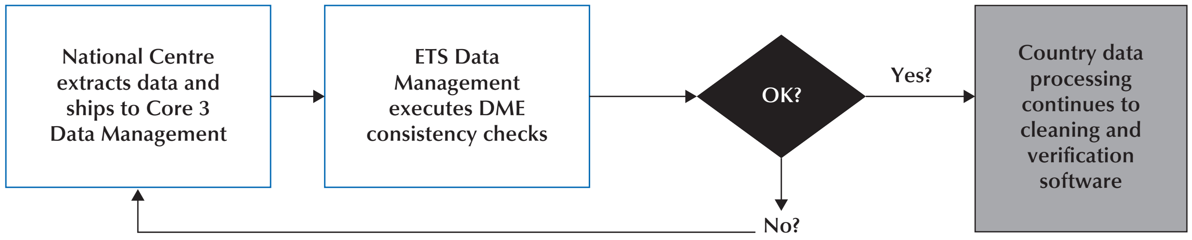
Pre-processing

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. Running the DME software was one of the first consistency 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 data quality.

These checks in addition of other internal checks for coding were executed and any inconsistencies were compiled into a report and returned 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 document 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 series of consistency checks to be sure all necessary issues are resolved and/or documented. In this initial step of processing, returning data inconsistencies to the National Centres was an iterative process with some times up to 4-5 iterations of data changes/updates from the country. Once resolved, the data continued to the next phase of the internal process – loading the database into the cleaning and verification software.

■ Figure 10.2 ■

Overview of the delivery and pre-processing phase


National Centre extracts data and ships to Core 3 Data Management
ETS Data Management executes DME consistency checks
OK?
Yes?
No?
Country data processing continues to cleaning and verification software

Initial database load into SQL server and the cleaning and verification software

With the 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 and used a processing software that merged datasets in SAS, but had the ability to produce both SAS and SPSS datasets. During processing, one to two analysts independently cleaned country databases, focusing on one country at a time in order to complete all necessary phases of quality assurance, in order to produce both SAS and SPSS datasets to the country and other contractors.

The first step in this process was to load the DME database onto the ETS Data Management cleaning and verification server. 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 2015, patch files were released by ETS

Data Management and applied to the SQL database by the National Data Manager to correct errors in the codebook or to modify the consistency checks in the DME software. For example, a patch may be issued if an item was misclassified as having 4 category response options instead of 5.

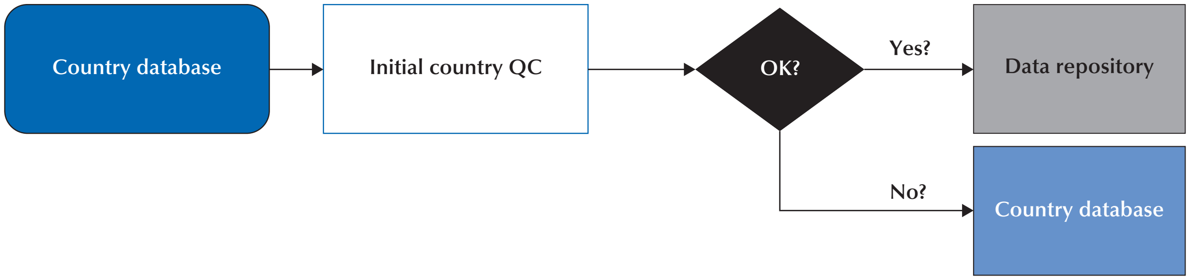
- The country database had the correct profile as dictated by the country's selected international options (e.g. Financial Literacy, UH booklet, etc.).
- The number of cases in the data files by country/language were in agreement with the sampling information collected by Westat.
- All values for variables 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.

Integration

After the initial load into the data repository and completion of early processing checks (Figure 10.3), the database entered the next phase of processing: Integration (Figure 10.4). 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 representing each of the respondent types: student, school, and teacher (where applicable). Additionally, parent questionnaire data was merged with their child/student data.

■ Figure 10.3 ■

Initial load of the National Centre database into SQL server for processing



In the main survey, the integration phase was a critical juncture because data management was able to analyse the data collected within the context of the sampling information supplied by the sampling contractor, Westat. Using this sampling information –captured in the Student Tracking Form – extensive quality control checks were applied to the data in this phase. Over 80 quality assurance checks were performed on the database during this phase, including specific checks such as: verifying student data discrepancies of students who are marked as present but do not have test or questionnaire data; students who are not of the "allowable" PISA age; and students who are marked absent but have valid test or questionnaire data. As a result of these quality assurance checks, a quality control report was generated and delivered to countries to resolve outstanding issues and inconsistencies. This report was referred to as the Quality Control ("Country QC") Report.

In this report, ETS Data Management 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):

- FORMCODE was blank or not valid.
- Student was missing key data needed for sampling and processing.
- Student was not in the allowable age.

- Student was not represented in the STF.
- Students who were marked absent yet had records.
- Mother's or father's occupation appeared invalid or needed clarification because it was not of length 4.
- Student's grade was lower than expected.
- On the Teacher Questionnaire, a teacher was marked as a "non-participant"[2], yet data existed.

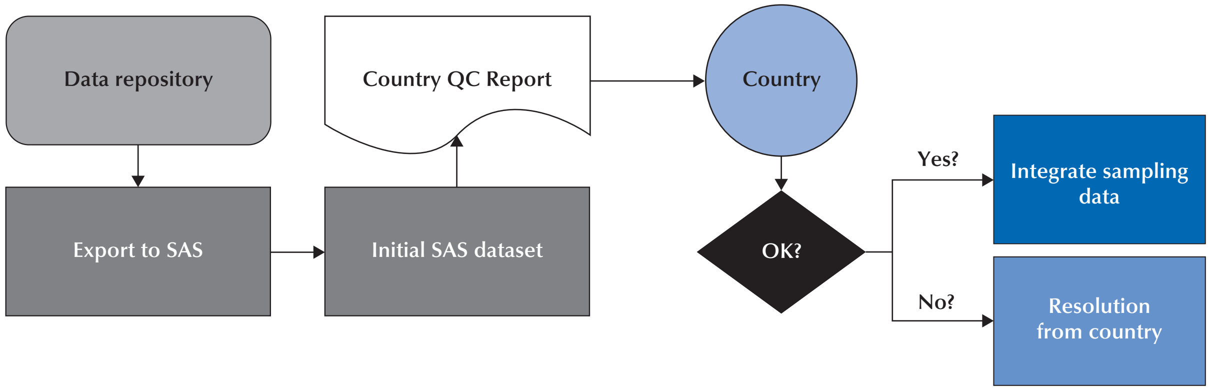
In addition to quality control reporting, a series of important data processing steps occurred during integration.

- **Item Cluster Analysis:** For the purposes of data processing, it is often convenient to disaggregate a single variable into a collection of variables. To this end, a respondent's single booklet number was interpreted 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 which detects any disconnect between the student delivery system and 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 always captured accurately. This was not necessarily true, however, in the case of computer-based administrations. While the student delivery system captures a student's response, it fails to 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 were summed appropriately to give timing for entire sections of the assessment.
- **SDS Post-processing:** Necessary changes in the student delivery system (SDS) were sometimes detected after the platform was already in use. For example, a test item that was scored by the SDS may have had an error in the interpretation of a correct response, which was corrected in the SDS post-processing. These and other issues were resolved by the SDS developers and new scored response data was processed, issued, and merged by the Data Management team.

Following the Integration phase of data processing, the Country QC reports were generated and distributed to the National Centres. National Project Managers were asked to review the report and to address any reported violations. National Centres corrected or verified inconsistencies in the database from this report and returned the revised database to the Data Management contractor within a specific timeframe. Additionally, all data revisions were documented directly in the Country QC report for delivery to Data Management. After receiving the revised database, the Data Management team repeated the pre-processing phase to ensure no new errors were reported and, if no errors were found, the Data Management team re-executed the integration step. As with the pre-processing consistency checks phase, the integration step 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 team in order to resolve issues. After all checks were revised and documented by the National Centre and no critical data violations remained, the data moved to the next phase in processing – i.e. national adaptation harmonisation.

■ Figure 10.4 ■

Integration process overview



HARMONISATION

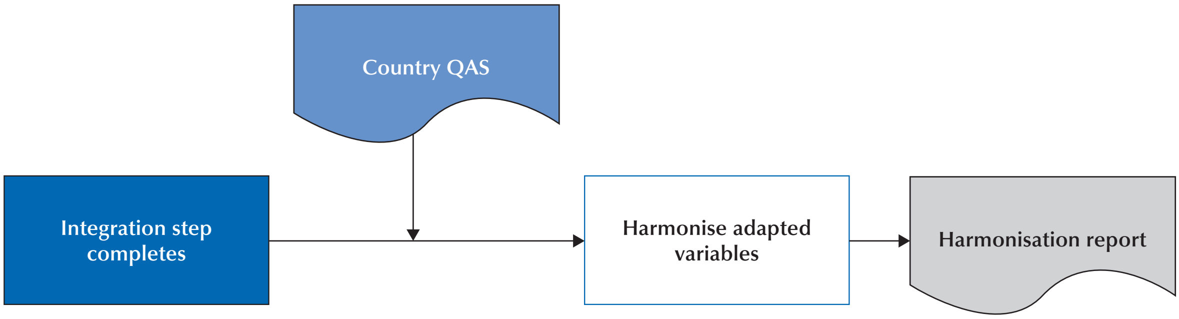
Overview of the workflow

As mentioned earlier in this chapter, 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, referred to as "national adaptations". As a result, changes to variables proposed by a National Centre occurred during the translation and adaptation process. National adaptations for questionnaire variables were agreed upon by the Background Questionnaire contractors. These discussions regarding adaptations happened in the "negotiation" phase between the country and the contractor as well as the translation verification contractor. 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 national adaptation requiring harmonisation code. 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 BQ contractor, with the assistance of the translation verification contractor, to ensure the QAS was complete and reflected the country's intent and interpretation. Once adaptations were approved by the BQ contractor, countries were able to implement their approved national adaptations (using their QAS as a reference tool) in their questionnaire material. National Centres were required to document and implement all adaptations in the following resources: QAS and the DME.

Any issues surrounding the national adaptations were handled by the country as well as by both the BQ contractor and the Data Management contractor. Official BQ contractor approval of the harmonisation SAS code was required for data processing. Additionally, the BQ contractor was responsible for reviewing the harmonisation reports produced by ETS Data Management for any issues or concerns with national adaptations. The National Centres also reviewed these harmonisation reports and contacted both the BQ contractor and the Data Management contractor with any issues or changes. Changes were documented in the country QAS file. Following any change or modification, the data management team repeated the harmonisation stage in order to check the proposed changes.

■ Figure 10.5 ■

Harmonisation process overview



Harmonisation, or harmonised variables

In general, 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 with SAS code. This code was implemented into the data management cleaning and verification software in order to handle these harmonised variables during processing.

Additionally, harmonisation consisted of 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; or perhaps two questions were merged (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).

VALIDATION

After the harmonisation process, 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, ETS Data Management collaborated with the BQ contractor to develop a series of validation checks that were performed on the data following harmonisation. Validation checks are consistency checks that provide National Centres with more detail concerning extreme and/or inconsistent values in their data. These violations of the validation checks were displayed in a validation report, which was shared with countries and contactors to observe these inconsistencies and make improvements for the next cycle of PISA. In the PISA 2015 main survey, National Centres did not make changes to revise these extreme and/or inconsistent values in the report. Rather, National Centres were instructed to leave the data as it is and make recommendations for addressing these issues in the data collection process during the next cycle of PISA. Although data modifications were not made for many of these validation checks, ETS Data Management required National Centres to document and provide more information into the nature of these data inconsistencies. Generally, validation checks of this nature captured inconsistent student, school and teacher data. For example, these checks may capture an inconsistency between the total number of years teaching to the number of years teaching at a particular school (TE0001); 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 (ST059Q02TA). Throughout this PISA cycle, these validation checks often served as valuable feedback for data quality.

Treatment of inconsistent and extreme values in PISA 2015 main survey data

During the preparations for the main survey international database release, some National Centres raised the issue of how to handle some extreme and/or inconsistent values within the data. The Data Management contractor, the Background Questionnaire contractor and the OECD agreed on implementing a specific approach to manage the extreme and/or inconsistent values present within the data.

Concerning the special handling of these inconsistent and/or extreme values, the following principles were followed:

- Support the results of DME software consistency checks from the PISA 2015 main survey. 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).[3]
- Support the results of the validation checks from PISA 2015 main survey. In particular, it is key to note that cases that corresponded to selections from drop-down menus were not invalidated (for example, the variable, EC029Q01NA, from the Educational Career Questionnaire item, “How many years altogether have you attended additional instruction?”), however implausible.
- Apply stringent consistency and validity checks while computing derived variables.[4]

The specific range restriction rules for PISA 2015 are located in Figure 10.6 at the end of this chapter.

SCORING AND DERIVATION

After validation, the next phase of data management processing involved parallel processes that occur with test data and questionnaire data:

- Scoring of test responses captured in paper booklets.
- Derivation of new variables from questionnaires.

Scoring overview

The goal of the PISA assessment is to ensure comparability of results across countries. As a result, scoring for the tests 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. To aid in this process, the Data Management team 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 responses. The Data Management team was not only responsible for generating the SAS code to implement these values, 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 the data, the Data Management team 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 miscoded 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 were completed. If the reported data inconsistencies were resolved, the scoring process was complete and the data proceeded to the next phase of processing.

The scoring variables also served as a valuable quality control check. If any items appeared to function not as expected (too difficult or too easy), further investigation was carried out to determine if a booklet printing error occurred or if systematic errors were introduced during data entry.

Derived variables overview

The SAS derived variable code was generated by the BQ contractor, DIPF, for implementation into the Data Management cleaning and verification software at this step in the process. The derived variable code 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 in which it was outlined that approximately 219 derived variables were calculated from PISA MS data.

Further explained in the MS Analysis Plan, for all questions in the MS questionnaires that were not converted into derived variables, the international database contained item-level data as obtained from the delivery platform. These included single-item constructs that could be measured without any transformation (e.g., ST002 Study program, ST016 Life satisfaction, ST021 Age of immigration, ST111 ISCED-level of educational aspiration, SC013 School type: public vs private, SC014 School management), as well as multi-item questions that were used by analysts for their respective needs (e.g., ST063 School science courses attended, ST064 Freedom of curricular choice, ST076/078 Activities before/after school, and most questions from the School Questionnaire). Derived variables were specified in line with previous cycles of PISA wherever possible. In terms of this alignment, first priority was given to alignment with PISA 2006, to enable comparison on science-related issues. Second priority was given to PISA 2012, to enable stability across recent and future cycles. For IRT scales, only alignment with PISA 2006 was included. See Chapter 16 for more information on derived variables.

As this phase of the processing was completed, all derivations were checked by DIPF. Any updates or recoding made to the derived variable code were completed and documented and redelivered to the Data Management team for use in the cleaning and verification software. Data files were refreshed appropriately with this new code to include all updates to these variables.

DELIVERABLES

After all data processing steps were complete and all updates to the data were made by National Centres to resolve any issues or inconsistencies, the final phase of data processing included the creation of deliverable files for all core contractors 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 and verification process contained critical applications to the data – such as the application of proxy scores, plausible values, background questionnaire scales, and weights. The dynamic feature of the cleaning and verification software allowed for the Data Management tea to tailor specific deliverables.

ETS Data Management produced a database containing the PISA 2015 data for National Centres and provided specific deliverables for core contractors as well as the OECD Secretariat according to particular specifications. In order to produce these customised files for contractors, 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

In order to prepare for the public release of PISA 2015 main survey data, ETS Data Management 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 proxy proficiency scores

for analysis. These data were later updated to include 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.

Files prepared for national centre data reviews

During the PISA 2015 main survey, the following files were prepared and released to National Centres at different stages of the data review:

- **Student combined data file** contained all student responses for test items (raw and scored), background questionnaire items, financial literacy items (if applicable), collaborative problem-solving items (if applicable), and optional questionnaire items such as Parent Questionnaire, Educational Career (EC) Questionnaire, Information and Computer Technology Literacy Familiarity (ICT) Questionnaire. These files included all raw variables, questionnaire indices, sampling weights, replicate weights, and plausible values.
- **School data file** contained all data from the School Questionnaires. These files included all raw variables, questionnaire indices, sampling weights, replicate weights, and plausible values.
- **Teacher data file** (if applicable) comprised data from the Teacher Questionnaire. These files included all raw variables, questionnaire indices and plausible values. In PISA 2015, Westat sampling did not calculate teacher weights and as such, there were no teacher weights in the data files.
- **Masked international database file** was a concatenated file of all countries provided further information for analysis to National Centres. In order to preserve country anonymity in this file, data files were 'masked' following specific guidelines from the OECD Secretariat that included issuing 'alternate' codes or required special handling for country identifiers.
- **Preliminary Public Use File** was produced toward the end of the PISA 2015 main survey and provided the National Centre with their country's own data as it would be presented in the final public release. These data included all country-requested variable suppressions. More information on the suppression period is discussed later in this chapter.
- **Analysis Reports** were delivered by data management and analysis and used by contractors and National Centres for quality control and validation purposes: plausibility of 1) distributions of background characteristics and 2) performance results for groups, especially in the in the extent to which they agree with expectations or external/historical information. These reports included:
  - **BQ Crosstabs:** An Excel file with cross tabulations of numeric categorical variables from the country's Background Questionnaire.
  - **BQ MSIGS:** An Excel file of summary statistics for all numerical variables from the country's Background Questionnaire.
  - **BQ SDTs:** Sets of country files 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.
  - **Item Analysis Reports:** The item analysis tables contained summary information about the response types given by the respondents to the cognitive items. They contained, for each country, the percent of individuals choosing each option for multiple-choice items or the percent of individuals receiving each score in the scoring guide for the constructed-response items. They also contained the international average percentages for each response category.

Records included in and excluded from the database

The following records were included in the database:

- student files
  - all PISA student respondents who participated in either the paper-based or computer-based assessment
  - all PISA students who had any response data or who were part of the original country sample
- school files
  - all participating schools – specifically, any school with a student included in the PISA sample and with a record in the school-level international database regardless of whether or not the school returned the School Questionnaire
- Teacher files
  - all PISA teacher participants that were included in the original sample.

Records excluded from the database

The following records were excluded the database[5]:

- student files
  - additional data collected by countries as part of national options
  - students who did not have the minimum response data to be considered a "respondent"[6]
  - students who refused to participate in the assessment sessions
- school files
  - additional data collected by countries as part of national options
- teacher files
  - teachers who refused to participate in the questionnaire.

Categorising missing data

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

1. Missing/blank is used to indicate the respondent was not presented the question according to the survey design or ended the assessment early and did not see the question. In the questionnaire data, it is only used to indicate that the respondent ended the assessment early or despite the opportunity, did not take the questionnaire.
2. No response/Omit indicates that the respondent had an opportunity to answer the question but did not respond.
3. Invalid is used to indicate a questionnaire item was suppressed by country request or that an answer was not conforming to the expected response. For a paper-based questionnaire, the respondent indicated more than one choice for an exclusive-choice question. For a computer-based questionnaire, the response was not in an acceptable range of responses, e.g., the response to a question asking for a percentage was greater than 100.
4. Not applicable indicates that a response was provided even though the response to an earlier question should have directed the respondent to skip that question, or the response could not be determined due to a printing problem or torn booklet. In the questionnaire data, it is also used to indicate missing by design (i.e. the respondent was never given the opportunity to see this question).
5. Valid skip indicates that the question was not answered because a response to an earlier question directed the respondent to skip the question. This code was assigned during data processing.

Data management and confidentiality, variable suppressions

During the PISA 2015 cycle, some country regulations and laws restricted the sharing of data, as originally collected, 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 clearly impacts the analytical utility of the data. Therefore, both goals must be carefully balanced. As a general directive for PISA 2015, 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 in order 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, as well as technical errors in the data that could not be resolved through contractor cleaning and verification procedures. A listing of suppressions at the country variable-level is located in Figure 10.7 at the end of this chapter.

■ Figure 10.6 [Part 1/3] ■

PISA 2015 range restriction rules for inconsistent and extreme values for main survey data

Sequence	Dataset (STU, SCH, TCH)	Description of rule	SAS code
Student dataset			
1	STU	Invalidate if number of class periods per week in test language (ST059Q01TA) is greater than 40.	if (ST059Q01TA > 40) then ST059Q01TA =.I;
2	STU	Invalidate if number of class periods per week in maths (ST059Q02TA) is greater than 40.	if (ST059Q02TA > 40) then ST059Q02TA =.I;
3	STU	Invalidate if number of class periods per week in science (ST059Q03TA) is greater than 40.	if (ST059Q03TA > 40) then ST059Q03TA =.I;
4	STU	Invalidate if number of total class periods in a week (ST060Q01NA) is greater than 120.	if (ST060Q01NA > 120) then ST060Q01NA =.I;
5	STU	Invalidate if average number of minutes in a class period (ST061Q01NA) is less than 10 or greater than 120.	if (ST061Q01NA > 120 or ST061Q01NA < 10) then ST061Q01NA =.I;
6	STU	Invalidate if age of child starting ISCED 1 (PA014Q01NA) is greater than 14.	if PA014Q01NA > 14 then PA014Q01NA =.I;
7	STU	Invalidate if repeated a grade in ISCED3 (ST127Q03TA) but currently in ISCED2.	if ISCEDL = 2 then ST127Q03TA =.I;
8	STU	Mark as missing if learning time per week in maths (MMINS) is greater than 2400 min (40 hours).	if MMINS > 2400 then MMINS =.M;
9	STU	Mark as missing if learning time per week in test language (LMINS) is greater than 2400 min (40 hours).	if LMINS > 2400 then LMINS =.M;
10	STU	Mark as missing if learning time per week in science (SMINS) is greater than 2400 min (40 hours).	if SMINS > 2400 then SMINS =.M;

11	STU	Mark as missing if learning time per week in total (TMINS) is greater than 3000 min (50 hours) or less than the sum of the parts (MMINS, LMINS, SMINS).	if TMINS > 3000 then TMINS =.M;
			if TMINS < sum(LMINS, MMINS, SMINS) then TMINS =.M;

12	STU	Mark as missing if out-of-school study time per week (OUTHOURS) is greater than 70 hours.	if OUTHOURS > 70 then OUTHOURS = .M;
13	STU	Invalidate if age started ISCED 1 is greater than 16 or less than 2.	if (ST126Q02TA > 16 or ST126Q02TA < 2) then ST126Q02TA =.I;
School dataset			
1	SCH	Invalidate if number of computers connected to the internet (SC004Q03TA) is greater than the number of computers available to students (SC004Q02TA).	if SC004Q03TA > SC004Q02TA then SC004Q03TA =.I;
2	SCH	Invalidate if number of portable computers (SC004Q04NA) is greater than the number of computers available to students (SC004Q02TA).	if SC004Q04NA > SC004Q02TA then SC004Q04NA =.I;
3	SCH	Invalidate if total number of full time teachers (SC018Q01TA01) is negative.	if (SC018Q01TA01 < 0) then SC018Q01TA01 =.I;
4	SCH	Invalidate if number of full time certified teachers (SC018Q02TA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q02TA01 > SC018Q01TA01 then SC018Q02TA01 =.I;
5	SCH	Invalidate if number of full time Bachelor degree teachers (SC018Q05NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q05NA01 > SC018Q01TA01 then SC018Q05NA01 =.I;
6	SCH	Invalidate if number of full time Master's degree teachers (SC018Q06NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q06NA01 > SC018Q01TA01 then SC018Q06NA01 =.I;
7	SCH	Invalidate if number of full time ISCED 6 teachers (SC018Q07NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q07NA01 > SC018Q01TA01 then SC018Q07NA01 =.I;
8	SCH	Invalidate if number of part time certified teachers (SC018Q02TA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q02TA02 > SC018Q01TA02 then SC018Q02TA02 =.I;
9	SCH	Invalidate if number of part time Bachelor degree teachers (SC018Q05NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q05NA02 > SC018Q01TA02 then SC018Q05NA02 =.I;
10	SCH	Invalidate if number of part time Master's degree teachers (SC018Q06NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q06NA02 > SC018Q01TA02 then SC018Q06NA02 =.I;
11	SCH	Invalidate if number of part time ISCED 6 teachers (SC018Q07NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q07NA02 > SC018Q01TA02 then SC018Q07NA02 =.I;
12	SCH	Invalidate if total number of full time science teachers (SC019Q01NA01) is negative.	if (SC019Q01NA01 < 0) then SC019Q01NA01 =.I;
13	SCH	Invalidate if number of full time science teachers (SC019Q01NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC019Q01NA01 > SC018Q01TA01 then SC019Q01NA01 =.I;

■ Figure 10.6 [Part 2/3] ■

PISA 2015 range restriction rules for inconsistent and extreme values for main survey data

Sequence	Dataset (STU, SCH, TCH)	Description of rule	SAS code
14	SCH	Invalidate if number of full time certified science teachers (SC019Q02NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC019Q02NA01 > SC018Q01TA01 then SC019Q02NA01 =.I;
15	SCH	Invalidate if number of full time ISCED 5A science teachers (SC019Q03NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC019Q03NA01 > SC018Q01TA01 then SC019Q03NA01 =.I;
16	SCH	Invalidate if number of part time science teachers (SC019Q01NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC019Q01NA02 > SC018Q01TA02 then SC019Q01NA02 =.I;
17	SCH	Invalidate if number of part time certified science teachers (SC019Q02NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC019Q02NA02 > SC018Q01TA02 then SC019Q02NA02 =.I;
18	SCH	Invalidate if number of part time ISCED 5A science teachers (SC019Q03NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC019Q03NA02 > SC018Q01TA02 then SC019Q03NA02 =.I;
19	SCH	Invalidate if number of full time certified science teachers (SC019Q02NA01) exceeds total number of full time science teachers (SC019Q01NA01).	if SC019Q02NA01 > SC019Q01NA01 then SC019Q02NA01 =.I;
20	SCH	Invalidate if number of full time ISCED 5A science teachers (SC019Q03NA01) exceeds total number of full time science teachers (SC019Q01NA01).	if SC019Q03NA01 > SC019Q01NA01 then SC019Q03NA01 =.I;
21	SCH	Invalidate if number of part time certified science teachers (SC019Q02NA02) exceeds total number of part time science teachers (SC019Q01NA02).	if SC019Q02NA02 > SC019Q01NA02 then SC019Q02NA02 =.I;
22	SCH	Invalidate if number of part time ISCED 5A science teachers (SC019Q03NA02) exceeds total number of part time science teachers (SC019Q01NA02).	if SC019Q03NA02 > SC019Q01NA02 then SC019Q03NA02 =.I;
23	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; end;
24	SCH	Invalidate if percentage of teaching staff (SC025Q01NA) is greater than 100%.	if SC025Q01NA > 100 then SC025Q01NA =.I;
25	SCH	Invalidate if percentage of science teacher staff (SC025Q02NA) is greater than 100%.	if SC025Q02NA > 100 then SC025Q02NA =.I;
26	SCH	Invalidate if percentage of students with <heritage language> different than <test language> (SC048Q01NA) is greater than 100%.	if SC048Q01NA > 100 then SC048Q01NA =.I;
27	SCH	Invalidate if percentage of students with special needs (SC048Q02NA) is greater than 100%.	if SC048Q02NA > 100 then SC048Q02NA =.I;
28	SCH	Invalidate if percentage of students from disadvantaged homes (SC048Q03NA) is greater than 100%.	if SC048Q03NA > 100 then SC048Q03NA =.I;
29	SCH	Invalidate if percentage of parents that initiated discussion on child (SC064Q01TA) is greater than 100%.	if SC064Q01TA > 100 then SC064Q01TA =.I;
30	SCH	Invalidate if percentage of parents where teacher initiated discussion on child (SC064Q02TA) is greater than 100%.	if SC064Q02TA > 100 then SC064Q02TA =.I;
31	SCH	Invalidate if percentage of parents participated in school government (SC064Q03TA) is greater than 100%.	if SC064Q03TA > 100 then SC064Q03TA =.I;
32	SCH	Invalidate if percentage of parents that volunteered in extracurricular activities (SC064Q04NA) is greater than 100%.	if SC064Q04NA > 100 then SC064Q04NA =.I;
33	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;
34	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;
35	SCH	Invalidate if total number of part time teachers (SC018Q01TA02) is negative.	if SC018Q01TA02 < 0 then SC018Q01TA02 =.I;
36	SCH	Mark index of computer availability (RATCMP1) as missing if there are only 10 or less students in the modal grade.	If SC004Q01TA <= 10 then RATCMP1 =.M;
37	SCH	Mark index of computers connected to the Internet (RATCMP2) as missing if there are only 10 or less students in the modal grade.	If SC004Q01TA <= 10 then RATCMP2 =.M;
38	SCH	Recode student-teacher ratio (STRATIO) to set the minimum number of teachers at 1 and then to set the final ratio to a maximum of 100 and a minimum of 1.	if nmiss(SCHSIZE,TOTAT) = 0 then STRATIO = max(min(SCHSIZE/max(1, TOTAT), 100), 1); else STRATIO =.M;

■ Figure 10.6 [Part 3/3] ■

PISA 2015 range restriction rules for inconsistent and extreme values for main survey data

Sequence	Dataset (STU, SCH, TCH)	Description of rule	SAS code
Teacher dataset			
1	TCH	Invalidate if number of years teaching at school (TC007Q01NA) exceeds reported age (TC002Q01NA) minus 15.	if TC007Q01NA > (TC002Q01NA – 15) then TC007Q01NA =.I;
2	TCH	Invalidate if total number of years teaching (TC007Q02NA) exceeds reported age (TC002Q01NA) minus 15.	if TC007Q02NA > (TC002Q01NA – 15) then TC007Q02NA =.I;
3	TCH	Invalidate if years working as a teacher in total (TC007Q02NA) is less than years working as a teacher in this school (TC007Q01NA).	if TC007Q01NA > TC007Q02NA then TC007Q01NA =.I;
4	TCH	Invalidate if proportion of teacher education dedicated to <broad science> and technology content (TC029Q01NA) + <school science> (TC029Q02NA) + general topics (TC029Q03NA) + other topics (TC029Q04NA) is less than 98% or greater than 102%.	if sum(TC029Q01NA, TC029Q02NA, TC029Q03NA, TC029Q04NA) > 102 or sum(TC029Q01NA, TC029Q02NA, TC029Q03NA, TC029Q04NA) < 98 then do; TC029Q01NA =.I; ,TC029Q02NA =.I; TC029Q03NA =.I; TC029Q04NA =.I; end;
5	TCH	Invalidate if proportion of professional development activities dedicated to <broad science > and technology content (TC030Q01NA) + <school science> (TC030Q02NA) + general topics (TC030Q03NA) + other topics (TC030Q04NA) is less than 98% or greater than 102%.	if sum(TC030Q01NA, TC030Q02NA, TC030Q03NA, TC030Q04NA) > 102 or sum(TC030Q01NA, TC030Q02NA, TC030Q03NA, TC030Q04NA) < 98 then do; TC030Q01NA =.I; TC030Q02NA =.I; TC030Q03NA =.I; TC030Q04NA =.I; end;

■ Figure 10.7 [Part 1/2] ■

PISA 2015 main survey country/variable suppression list

Country	Variable
AUT	Stratum SC002Q01TA, SC002Q02TA, SCHSIZE
AUS	Student financial literacy data
BEL (Flemish only)	SC013Q01TA, SC014Q01NA, SC016Q01TA, SC016Q02TA, SC016Q03TA, SC016Q04TA, SCHLTYPE
CHN	Stratum, Region
DEU	STRATUM
ISR	SC013, SC014, SC016, SCHLTYPE, STRATUM
ITA	STRATUM
QCY[1]	STRATUM, LANGTEST_COG, LANGTEST_QQQ, LANGTEST, SC001Q01TA
KAZ	STRATUM
NZL	SC002Q01TA, SC002Q02TA, SC004Q01TA, SC004Q02TA, SC014Q01NA, SCHSIZE
PRI, QNC, QMA[2]	All school variables, All teacher variables, CNTSCHID[3], ST001D01T, ST003D02T, ST003D03T, ST005Q01TA, ST006Q01TA, ST006Q02TA, ST006Q03TA, ST006Q04TA, ST007Q01TA, ST008Q01TA, ST008Q02TA, ST008Q03TA, ST008Q04TA, ST019AQ01T, ST019BQ01T, ST019CQ01T, ST021Q01TA, ST022Q01TA, AGE, ISCEDL, ISCEDD, ISCEDO, GRADE, IMMIG, MISCED, FISCED, HISCED, BMMJ1, BFMJ2, HISEI, PARED, COBN_F, COBN_M, COBN_S, LANGN, OCOD1, OCOD2, UNIT, WVARSTRR
SVN	ST063, ST064, ST065, ST103, ST104, ST107, TDTEACH, PERFEED, ADINST
SWE	ST003D02T, ST003D03T SC001Q01TA, SC002Q01TA, SC002Q02TA, SC003Q01TA, SC004Q01TA, SC013Q01TA, SC014Q01NA, SC016Q01TA, SC016Q02TA, SC016Q03TA, SC016Q04TA, SC018Q01TA01, SC018Q01TA02, SC018Q02TA01 SC018Q02TA02, SC018Q05NA01 SC018Q05NA02, SC018Q06NA01 SC018Q06NA02, SC018Q07NA01 SC018Q07NA02, SC019Q01NA01 SC019Q01NA02, SC019Q02NA01 SC019Q02NA02, SC019Q03NA01 SC019Q03NA02, SC048Q01NA SC048Q02NA, SC048Q03NA

■ Figure 10.7 [Part 2/2] ■

PISA 2015 main survey country/variable suppression list

Country	Variable
TAP	STRATUM
THA	STRATUM

1. QCY data is suppressed from the public use files. Variables were suppressed in the national data files.
2. QNC and QMA are the United States state samples analyzed in the PISA 2015 main survey.
3. With this suppression request, all school and teacher data is suppressed. As a result, CNTSCHID is suppressed in all data files.

Notes

1. "Data Management" refers to the collective set of activities and tasks that each country had to perform to produce the required national database.

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

3. For example, if an inconsistency existed between age and seniority, the proposed rules invalidates seniority but keeps "age".

4. With this principle, the original values were kept, while the values for the derived variable may have the applied "invalid" rule.

5. Due to issues identified during data adjudication, data from Argentina, Kazakhstan, Malaysia and Albania, student questionnaire data (only) have been extracted into a separate file for analysis.

6. To be considered a "respondent" the student must have one test item response and a minimum number of responses to the student background questionnaire (that included responses for ST012 or ST013); or, responded to at least half of the number of test items in his or her booklet/form.