Chapter 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 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 2018 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. 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 individual 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 2018:

- standards, guidelines, and recommendations for data management within countries
- data management software, manuals, codebooks, and training videos 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 organizations for information and consultation:

- ETS (Project Management - Core A): 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 (Background Questionnaires - Core A): 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 C): 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 A): Key to the implementation of the PISA assessment in countries, Westat's Survey Operations team supported countries through the PISA 2018 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 their 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.

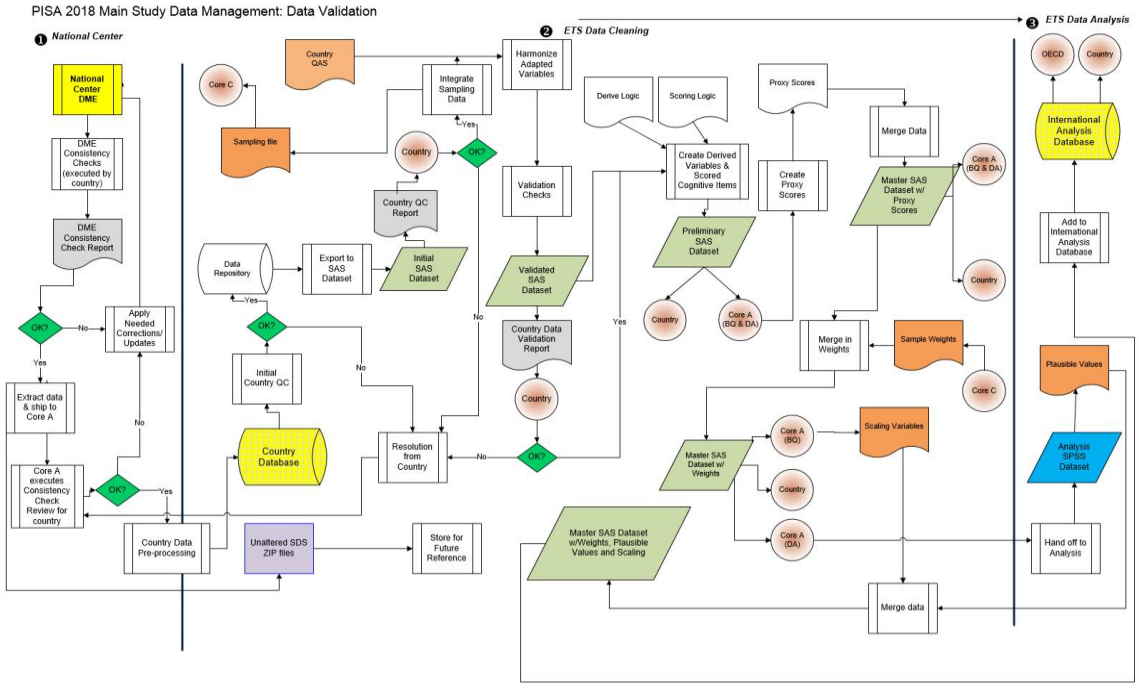
Data management at the national level

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. "Data Management" refers to the collective set of activities and tasks that each country had to perform to produce the required national database. 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, the following: 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 (DME). Figure 10.1 provides the workflow of the data management process for PISA 2018.

Figure 10.1 : 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 2018 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 presented a challenge and an opportunity to develop a system that both accepted and processed these electronic data and their variety of formats and supported 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 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 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 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. Each national codebook is a copy of the international codebook where the datasets corresponding to national options not implemented in the country have been hidden. For example, all codebooks in countries participating in the PBA 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) added to the appropriate datasets. The CBA codebooks are also tested using data obtained from the country's testing of their platform.

The codebook is delivered to each country as a read-only "template" file, which the DME application will copy into an active database file. The NDM must then confirm that the template file will create a codebook and the codebook will generate and support the appropriate datasets for their national options. The CBA countries are then requested to import their test data to ensure that the national adaptations and additions are properly handled. The PBA countries are first required to add the variables for their own national adaptations and additions to the questionnaire datasets, as there were not QAT documents available for these countries. They are then required to test the manual entry of the questionnaire data to confirm that the national variables are properly presented, and in their correct sequence. Similarly, those CBA countries who elect the Parent Questionnaire 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 for review and troubleshooting for any issues that may arise during data processing.

The DME application permits three levels of password-controlled access to the database. 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.

The DME application supports three modes of inputting data into the database: 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 linear form, such as a questionnaire, or 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 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 Excel and CSV files 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 PISA Imports category includes 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 the KeyQuest 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 collect and enter 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 getting 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:

- check for logical 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.
- check for logical consistency between the cognitive response data files and their corresponding OECS datasets to ensure that all respondents received codes for the open-ended items.
- provide counts of certain aspects of the database, such as number of students by language of survey.
- list the contents of certain 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 reports of the first two types and either resolve the noted discrepancies or provide an explanation for why they could not be resolved. 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 with all variables designated for suppression (e.g. Personally Identifiable Information) 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

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 the quality of the data received.

These checks, in addition to other internal checks for coding, were executed upon receipt of the data, 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 highlighted issues are resolved and/or documented. In this initial step of processing, returning data inconsistencies to the National Centres was an iterative process with sometimes 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.

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 also had the ability to produce both SAS and SPSS datasets. During processing, one or two analysts independently cleaned country databases, focusing on one country at a time in order to complete all necessary phases of quality assurance. The end goal was to produce both SAS and SPSS datasets to be delivered back to the country, and other contractors.

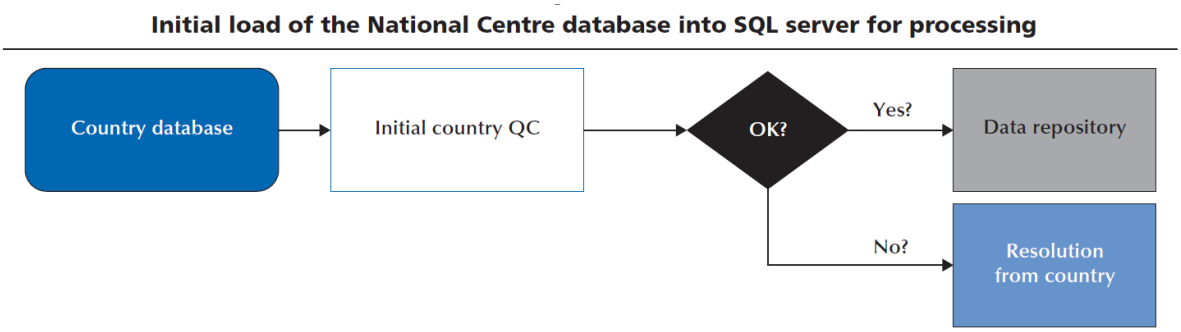
The first step in this process was to load the DME database onto the ETS Data Management cleaning and verification server. With the initial load of the database, specific quality assurance checks were applied to the data. These checks ensured:

- 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 2018, patch files were released by ETS Data Management and applied to the SQL 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, UH booklet, 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.

Integration

After the initial load into the data repository and completion of early processing checks (see Figure 10.2), the database entered the next phase of processing: Integration (see Figure 10.3). 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.

Figure 10.2: Initial load of the National Centre database into SQL server for processing



During data processing, the integration phase was critical because data management 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 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 target 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):

- Test 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 Student Tracking Form
- Students who were marked absent yet had records
- Student's grade was lower than expected
- On the Teacher Questionnaire, a teacher was marked as "non-participant" (absent, excluded, or refused to participate in the session), yet data existed for that teacher.

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 to detect any disconnect between the student delivery system and

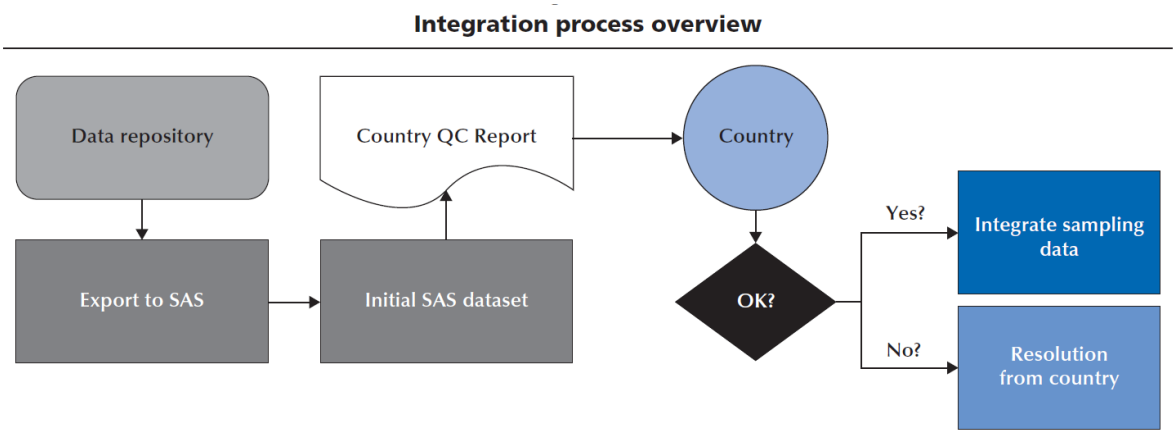
the sampling design. Any discrepancies discovered were resolved by contacting the appropriate contractors.

- 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 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 Core A Data Management.

Following the Integration phase of data processing, the Country Quality Control reports were generated and distributed to the National Centres. National Project Managers were asked to review the report and to address any reported issues. 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 issues or 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 10.3: 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 or contexts. These adaptations are referred to as "national adaptations." As a result, changes to variables by a National Centre were proposed 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. When necessary, 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.

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. Issues detected by with the validation 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. In the PISA 2018 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. Generally, validation checks captured inconsistent student, school and teacher data. For example, these checks may capture 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 2018 main survey data

Following the approach implemented in PISA 2015 regarding 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.

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

- Support the results of DME software consistency checks from the PISA 2018 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). For example, if an inconsistency existed between age and seniority, the proposed rules invalidates seniority but keeps "age".
- Apply stringent consistency and validity checks while computing derived variables. With this principle, the original values were kept, while the values for the derived variable may have the applied "invalid" rule.

The specific range restriction rules for PISA 2018 are presented in Tables 10.1 to 10.3 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 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. 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 each response. ETS Data Management was not only responsible for generating the SAS code 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, ETS Data Management 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 deemed 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

Code in SAS to create derived variables was generated by the BQ contractor, DIPF, for implementation into the Data Management cleaning and verification software 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.

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, as well as multi-item questions that were used by analysts for their respective needs. Whenever possible, derived variables were specified consistent with previous cycles of PISA. In terms of this alignment, the first choice was alignment with PISA 2009, to enable comparison on reading-related variables. Second choice was alignment with PISA 2015. This aimed to strike a balance and stability across recent and future cycles.

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 additions to the data – such as the addition of proxy scores, plausible values, background questionnaire scales, and sampling weights. The dynamic feature of the cleaning and verification software allowed for the Data Management team to generate customized files for delivery.

Core A Data Management produced a database containing the PISA 2018 data for National Centres and provided customized deliverables for core contractors as well as the OECD Secretariat. Each of these according to customized 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 2018 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.

The following data files were delivered:

- The Student combined data file contained all student responses for test items (raw and scored), background questionnaire items, global competence items, 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.
- The School data file contained all data collected with the School Questionnaires. These files included all raw variables, questionnaire indices, and other derived variables.
- The Teacher data file contained data from the Teacher Questionnaire. These files included all raw variables, questionnaire indices and derived variables.
- The Financial literacy data file contained 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. In order 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 was produced toward the end of the PISA 2018 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 on the suppression period is discussed later in this chapter.

In addition to the data files, Analysis Reports were delivered by data management and analysis and used by contractors and National Centres for quality control and validation purposes. In particular, these were 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: An Excel file with crosstabulations of categorical variables from the country's Background Questionnaire.
- BQ MSIGS: An Excel file of summary statistics for all continuous 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: Theses 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 respondents who participated in either the paper-based or computer-based assessment
- All respondents 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 the school returned the School Questionnaire

Teacher files

- All PISA teacher participants that were included in the original sample.

Financial literacy files

- Student respondents who: (1) took cognitive forms that had 1 hour of FL items and 1 hour of either Math or Reading (forms 73-84); and, (2) took cognitive forms that had 1 hour of Reading and 1 hour of Math (forms 1-12).

The following records were excluded from the database:

Student files

- Additional data collected by countries as part of national options contracts
- Students who did not have the minimum response data to be considered a "respondent". To be considered a "respondent" the student must have one test item response and a minimum number of responses to the student background questionnaire; or, responded to at least half of the number of test items in his or her booklet/form.
- 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 six different types of missing data:

- System Missing/Blank – used to indicate that the respondent was not presented the question according to the survey design or ended the questionnaire early and did not see the question.
- No Response – used to indicate the respondent had an opportunity to answer the question but did not respond. For derived variables, it is often used as an indicator for all different types of missing data.
- Invalid – used to indicate that the response was not appropriate or contradicted a prior response, e.g., the response to a question asking for a percentage was greater than 100.
- Not 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. In the cognitive data, it is used to indicate that the question was dropped/deleted during item calibration and not used during scaling.
- 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.
- 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 2018 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 2018, 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. A listing of suppressions at the country variable-level is in Table 10.4 at the end of this chapter.

Table 10.1: Range restriction rules for inconsistent and extreme values in the student file

Sequence	Description	SAS Code
1	Invalidate if number for an individual's weight is negative.	if (WB151Q01HA < 0) then WB151Q01HA=.I;
2	Invalidate if number for an individual's height is negative.	if (WB152Q01HA < 0) then WB152Q01HA=.I;
3	Invalidate if number of class periods per week in test language lessons (ST059Q01TA) is greater than 40.	if (ST059Q01TA > 40) then ST059Q01TA =.I;
4	Invalidate if number of class periods per week in maths (ST059Q02TA) is greater than 40.	if (ST059Q02TA > 40) then ST059Q02TA =.I;
5	Invalidate if number of class periods per week in science (ST059Q03TA) is greater than 40.	if (ST059Q03TA > 40) then ST059Q03TA =.I;
6	Invalidate if number of <class periods> per week in foreign language is greater than 40.	if (ST059Q04HA > 40) then ST059Q04HA= .I;
7	Invalidate if number of total class periods in a week (ST060Q01NA) is greater than 120 or less than 10	if (ST060Q01NA > 120 or ST060Q01NA < 10) and NOT MISSING(ST060Q01NA) then ST060Q01NA =.I;
8	Invalidate if average number of minutes in a class period (ST061Q01NA) is greater than 120 or less than 10.	if (ST061Q01NA > 120 or ST061Q01NA < 10) and NOT MISSING(ST061Q01NA) then ST061Q01NA =.I;
9	Invalidate if age of child starting ISCED 1 is greater than 16 or less than 2.	if (ST126Q01TA > 16 or ST126Q01TA < 2) and NOT MISSING(ST126Q01TA) then ST126Q01TA =.I;
10	Invalidate if age of child starting ISCED 1 (PA014Q01NA) is greater than 16 or less than 2.	if (PA014Q01NA > 16 or PA014Q01NA < 2) and NOT MISSING(PA014Q01NA) then PA014Q01NA =.I;
11	Invalidate if a child's ISCED level equals 2 and selects that he or she has repeated ISCED 2 or ISCED 3	if ISCEDL=2 and (ST127Q03TA=2 or ST127Q03TA=3) then ST127Q03TA =.I;

Table 10.2: Range restriction rules for inconsistent and extreme values in the school file

Sequence	Description	SAS Code
1	Invalidate if number of computers connected to the internet (SC004Q03TA) is greater than the number of computers available to students (SC004Q02TA).	if SC004Q03TA > SC004Q02TA and NOT MISSING(SC004Q02TA) then SC004Q03TA =.I;
2	Invalidate if number of portable computers (SC004Q04NA) is greater than the number of computers available to students (SC004Q02TA).	if SC004Q04NA > SC004Q02TA and NOT MISSING(SC004Q02TA) then SC004Q04NA =.I;
3	Invalidate if total number of full time teachers (SC018Q01TA01) is negative.	if (SC018Q01TA01 < 0) and NOT MISSING(SC018Q01TA01) then SC018Q01TA01 =.I;
4	Invalidate if number of full time certified teachers (SC018Q02TA01) is negative	if (SC018Q01TA02 < 0) and NOT MISSING(SC018Q01TA02) then SC018Q01TA02 =.I;
5	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;
6	Invalidate if number of full time Bachelor degree teachers (SC018Q05NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q05NA01 > SC018Q01TA01 and NOT MISSING(SC018Q01TA01) then SC018Q05NA01 =.I;
7	Invalidate if number of full time Master's degree teachers (SC018Q06NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q06NA01 > SC018Q01TA01 and NOT MISSING(SC018Q01TA01) then SC018Q06NA01 =.I;
8	Invalidate if number of full time ISCED 6 teachers (SC018Q07NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q07NA01 > SC018Q01TA01 and NOT MISSING(SC018Q01TA01) then SC018Q07NA01 =.I;
9	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;
10	Invalidate if number of part time Bachelor degree teachers (SC018Q05NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q05NA02 > SC018Q01TA02 and NOT MISSING(SC018Q01TA02) then SC018Q05NA02 =.I;
11	Invalidate if number of part time Master's degree teachers (SC018Q06NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q06NA02 > SC018Q01TA02 and NOT MISSING(SC018Q01TA02) then SC018Q06NA02 =.I;
12	Invalidate if number of part time ISCED 6 teachers (SC018Q07NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q07NA02 > SC018Q01TA02 and NOT MISSING(SC018Q01TA02) then SC018Q07NA02 =.I;
13	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;
14	Invalidate if percentage of teaching staff (SC025Q01NA) is greater than 100%.	if SC025Q01NA>100 then SC025Q01NA =.I;
15	Invalidate if percentage of science teacher staff (SC025Q02NA) is greater than 100%.	*if SC025Q02NA>100 then SC025Q02NA =.I;* no science teacher in 2018;
16	Invalidate if percentage of students with <heritage language> different than <test language> (SC048Q01NA) is greater than 100%.	if SC048Q01NA>100 then SC048Q01NA =.I;
17	Invalidate if percentage of students with special needs (SC048Q02NA) is greater than 100%.	if SC048Q02NA>100 then SC048Q02NA =.I;

Sequence	Description	SAS Code
18	Invalidate if percentage of students from disadvantaged homes (SC048Q03NA) is greater than 100%.	if SC048Q03NA>100 then SC048Q03NA =.I;
19	Invalidate if percentage of parents that initiated discussion on child (SC064Q01TA) is greater than 100%.	if SC064Q01TA>100 then SC064Q01TA =.I;
20	Invalidate if percentage of parents where teacher initiated discussion on child (SC064Q02TA) is greater than 100%.	if SC064Q02TA>100 then SC064Q02TA =.I;
21	Invalidate if percentage of parents participated in school government (SC064Q03TA) is greater than 100%.	if SC064Q03TA>100 then SC064Q03TA =.I;
22	Invalidate if percentage of parents that volunteered in extracurricular activities (SC064Q04NA) is greater than 100%.	if SC064Q04NA>100 then SC064Q04NA =.I;
23	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;
24	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;

Table 10.3: Range restriction rules for inconsistent and extreme values in the teacher file

Sequence	Description	SAS Code
1	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	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	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 and NOT MISSING(TC007Q02NA) then TC007Q01NA =.I;
4	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	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;

Table 10.4: PISA 2018 Main Survey Country/Variable Suppression List

Country	Variable(s)
AUS	Student financial literacy data
AUT	STRATUM, ST001D01T, SC002Q01TA, SC002Q02TA, SCHSIZE
CAN	SC002Q01TA, SC002Q02TA, SC018Q01TA01, SC018Q01TA02, SC018Q02TA01, SC018Q02TA02, SC018Q05NA01, SC018Q05NA02, SC018Q06NA01, SC018Q06NA02, SC018Q07NA01, SC018Q07NA02, SC003Q01TA, STRATIO, SCHSIZE, TOTAT, CLSIZE
DEU	STRATUM
ISR	STRATUM
ITA	STRATUM
JPN	IMMIG
JOR	STRATUM
NOR	SC001Q01TA, SC013Q01TA, SC016Q01TA, SC016Q02TA, SC016Q03TA, SC016Q04TA, SC002Q01TA, SC002Q02TA, SC048Q01NA, SC048Q02NA, SC004Q01TA, SC004Q02TA, SC004Q03TA, SC004Q04NA, SC004Q07NA, SCHLTYPE, PRIVATESCH, STRATIO, SCHSIZE, TOTAT, PROATCE, PROAT5AB, PROAT5AM, PROAT6, SC002Q01TA, SC002Q02TA, SC004Q01TA, SC004Q02TA, SC018Q01TA01, SC018Q01TA02, SC018Q02TA01, SC018Q02TA02, SC018Q05NA01, SC018Q05NA02, SC018Q06NA01, SC018Q06NA02, LANGTEST_COG, LANGTEST_QQQ, ST003D02T
NZL	SC002Q01TA, SC002Q02TA, SC004Q01TA, SC004Q02TA, SC018Q01TA01, SC018Q01TA02, SC018Q02TA01, SC018Q02TA02, SC018Q05NA01, SC018Q05NA02, SC018Q06NA01, SC018Q06NA02, SC018Q07NA01, SC018Q07NA02, SCHSIZE, TOTAT
QCY	STRATUM, LANGTEST_COG, LANGTEST_QQQ, SC001Q01TA
SGP	LANGN, OCOD1, OCOD2, BMMJ1, BFMJ2
SWE	CLSIZE, SCHLTYPE, TOTAT, SC001Q01TA, SC013Q01TA, SC016Q01TA, SC016Q02TA, SC016Q03TA, SC016Q04TA, SC002Q01TA, SC002Q02TA, SC048Q01NA, SC048Q02NA, SC048Q03NA, SC004Q01TA, SC018Q01TA01, SC018Q01TA02, SC018Q02TA01, SC018Q02TA02, SC018Q05NA01, SC018Q05NA02, SC018Q06NA01, SC018Q06NA02, SC018Q07NA01, SC018Q07NA02, SC003Q01TA
THA	STRATUM