Chapter 11: DATA MANAGEMENT PROCEDURES

INTRODUCTION

In PISA-D, as in any international survey, a set of data collection requirements guide the creation of an international database that allows for valid within- and cross-country comparisons and inferences to be made. PISA-D standards and requirements serve as guidelines for contractors of the Consortium. 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 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-D is an international study that includes countries with unique education systems and cultural contexts. The PISA-D 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, the Consortium has conducted a series of consultations with the national representatives of participating countries in order to reflect country expectations in accordance with PISA-D Technical Standards. During these consultations, the data coding of the national adaptations to the instruments was discussed to ensure consistent recoding in an international format.

An important part of the data collection and management cycle is to not only 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, or render them unusable. For example, certain deviations of the standard testing procedures are particularly likely to affect test performance (e.g. 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 (i.e. the PISA-D 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 must be generalised to suit cognitive test instruments and background questionnaire instruments used in each participating country. As a result, a suite of products is provided to countries, including a comprehensive Data Management Manual, training

sessions and a range of other materials. In particular, National Project Managers (NPMs) and National Data Managers (NDMs) are equipped with data management software designed to carry out, in a consistent way, data management tasks, prevent the 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 level. In accordance with the PISA Technical Standards, the implementation of PISA among contractors of the Consortium is led by the Educational Testing Service (ETS). With the guidance from the OECD, ETS implemented and/or monitored the following:

- standards, guidelines, and recommendations for data management within countries
- data management software, manuals and codebooks to 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 Consortium, OECD and the National Centres
- preparation of data products for dissemination to Consortium, National Centres, the OECD and the public.

Additionally, at the international level, ETS data management and analysis practices relied on the following groups and organisations for information and consultation:

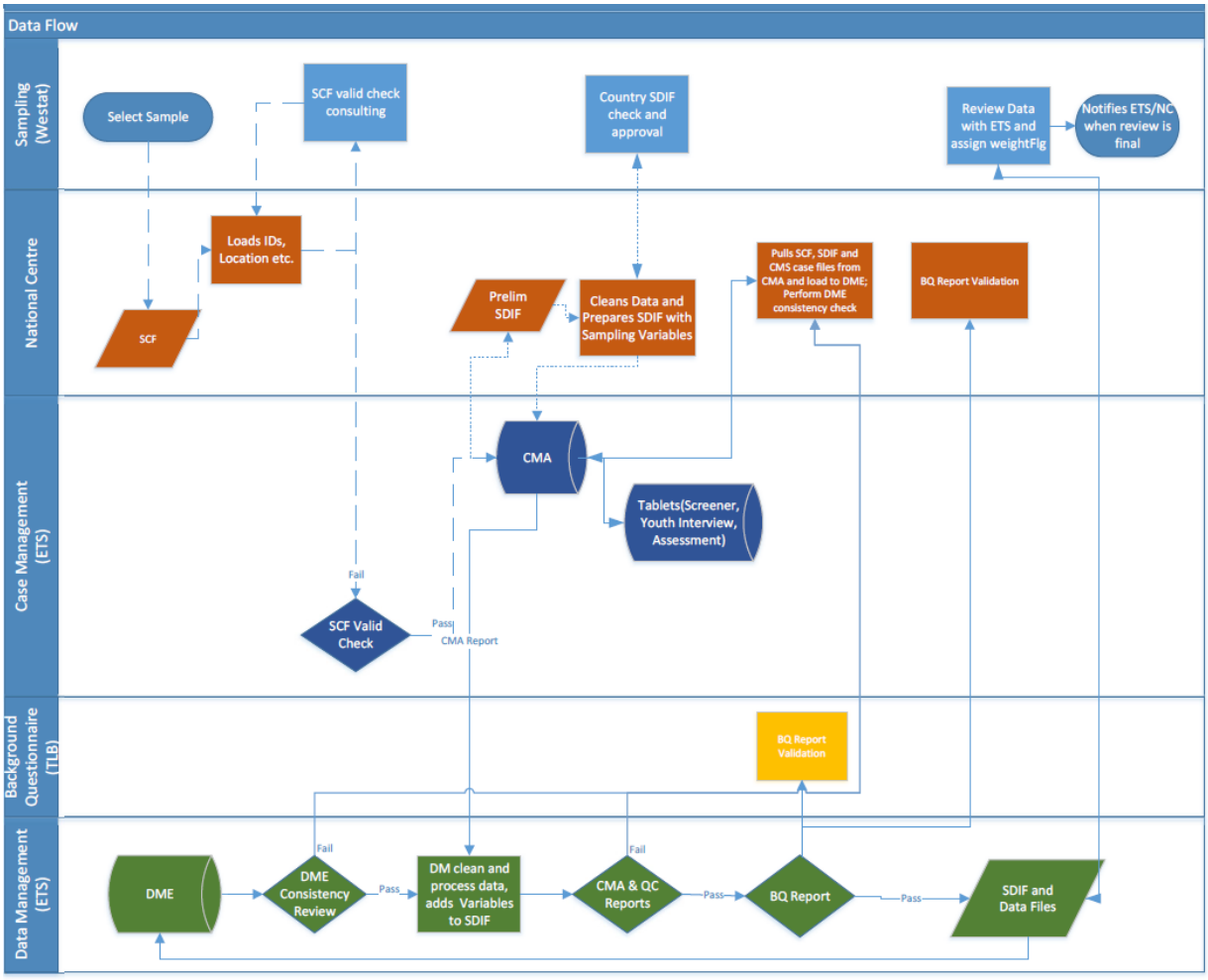
- ETS Project Management: ETS Project Management provided contractors with overview information on country details including timelines, testing dates, and support with country correspondence and deliverables planning.
- The Learning Bar (TLB): As the Background Questionnaire (BQ) experts, TLB provided BQ scaling and indices, BQ data, support for questionnaire negotiations with National Centres concerning questionnaire national adaptations, harmonisation and BQ derived variables.
- Westat (Sampling): Leading the Sampling tasks for PISA, Westat provided review and quality control support with respect to sampling and weighting. Westat was instrumental in providing guidance for quality assurance checks with regard to national samples.
- Westat (Survey Operations): Key to the implementation of the PISA assessment in countries, Westat's Survey Operations team supported countries through the PISA for Development cycle. Westat was responsible for specific quality assurance of the implementation of the assessment.
- 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.

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. Furthermore, as these data management tasks require technical skills for 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. The technical tasks for NDMs included collaboration with ETS on 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 concerns data entry and the execution of consistency checks from the primary data management software, the PISA Data Management Expert, or DME.

In PISA for Development, Figure 11.1 provides the workflow of the data management process at the national and international level.

Figure 11.1 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 the proper handling and generation of data. Additionally, more information is provided on the PISA for Development DME as well as the phases of the data management cleaning and verification process.

THE DATA MANAGEMENT PROCESS AND QUALITY CONTROL

In PISA, the collection of participant responses on a computer platform provided an opportunity for the accurate transcription of those responses and the collection of process data, including response actions and timing. This also presented a challenge to develop a system that accepted and processed these files and their variety of formats implemented in PISA-D. 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 Programme of International Assessment of Adult Competencies (PIAAC) and the most recent PISA cycle.

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 an international data codebook representing all the data to be collected by all countries without national adaptations. The national codebook for each country is simply a copy of the tested and approved international codebook. The National Data Manager (NDM) for each country is trained on and is responsible for implementing and testing the national adaptations to the delivered codebook.

The DME software provides 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 sampling data or the results of the occupational coding.

An important feature of the DME software is the ability to create multiple copies of the project codebook for use by entry operators 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 to ensure their completeness and accuracy. The first step was to use the DME software to create a consistency check report and review the results. National Centres were required to perform these checks frequently to highlight inconsistencies in sample status and missing instruments. They would then make appropriate adjustments or additions to the

database. They were also required to submit a consistency check report based on their submitted database, with all designated reports completely annotated with the reasons for all discrepancies found. The Data Management contractor also performed these consistency checks on receipt of the data as a quick and efficient way to verify data quality.

The contractor reviewed these reports and any outstanding inconsistencies were compiled into a report and returned to the National Centre with further information and/or corrections to the data. If necessary, National Centres resubmitted their data to the Data Management contractor after any changes were made to the database. Upon receipt of the redelivered database, the contractor refreshed the working database with the new data from the National Centre and restarted the pre-processing step, executing a new consistency check report to be sure all necessary issues were resolved and/or documented.

The resolution of data inconsistencies by the National Centres was an iterative process, with sometimes up to 4 or 5 iterations of data changes/updates from the country. Once these were resolved, the data were forwarded to the next phase of the data management process: loading the database into the cleaning and verification system.

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

When data preprocessing checks are complete, the country's database advances to the next phase of the process – data cleaning and verification. To reach the high-quality requirements of the PISA Technical Standards, the Data Management contractor created and used a processing software that merged datasets in SAS and had the ability to produce both SAS and SPSS datasets. During processing, Data Management analysts use this software to clean and verify the national database in order to produce both SAS and SPSS datasets for contractor analysis and National Centre review.

The first step in this process includes loading 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 includes all required patch files that 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 reports in the DME software. For example, a patch may be issued if an item was misclassified as a having 4 category response options instead of 5.
- the number of cases in the data files by Country/Language agree with the sampling information collected by Westat.
- all values for variables that used a value scheme are 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 may be a common occurrence with paper-based instruments (i.e. person most knowledgeable ("PMK")). Each suspicious data point was investigated and resolved by the country.
- response data that appeared to have no logical connection to other response data were validated to ensure correct IDs are captured.

With the completion of this initial processing step, a dataset is produced.

COGNITIVE PROCESSING

Cognitive data integration and quality control

After the initial load into the data repository and completion of early processing checks, the database enters the next phase: cognitive data processing. During this phase, data, which were structured within the country project database to assist in data collection, were restructured to facilitate data cleaning.

Cognitive processing includes the execution of a series of comprehensive quality assurance checks. As a result of these 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 and verified by the country.

In this report, ETS Data Management provided specific information to countries, including the name of the quality assurance check and the description of the check, as well as specific information, such as personal coordinates, for the cases that proved to be inconsistent or incorrect within the data. These checks were conducted in the following cases:

- participant was missing key data needed for sampling and processing
- participant was not in the allowable age
- participant was not represented in the SDIF
- participants who had valid CBA records, but no response data
- invalid or missing Youth Interview occupation codes.

In addition to quality control reporting, responses to items within a cluster were also evaluated during cognitive data processing. In order to verify the responses to items within clusters, the item responses for a participant were interpreted and coded into a single variable that represented the item clusters that were presented to the participant. An analysis was performed which detects any disconnect between the assessment and the sampling design. Any discrepancies discovered were resolved by contacting the appropriate Contractors.

As with the preprocessing consistency checks phase of data processing, the quality control report may involve several iterations of review if the National Centre does not appropriately address data inconsistencies. Frequently, one-on-one consultations were needed between the National Centre and the Data Management contractor in order to resolve issues.

Scoring analysis

The goal of the PISA-D assessment is to ensure comparability of results across countries. As a result, scoring for the cognitive assessment is a critical component of the data management processing. During this process, specific variables are created, and relevant student responses are inserted. To assist in this process, the Data Management contractor implemented rules from coding guides developed by the assessment development team. The coding guides organised in clusters, outline the value or score for responses. The Data Management contractor was not only responsible for generating the SAS code to implement these values, but was also responsible for implementing a series of validation checks on the data to determine any violations in scoring and/or any missing information. If any items appeared to function not as expected (too difficult or too easy), further investigation was carried out to determine if systematic errors were introduced during data entry.

Additionally, in this process, a scoring distribution was generated within cluster for all cognitive items for additional quality control. A summary report was produced that identified any potential issues within these scoring distributions as well as items that may be missing scoring information. When missing scores were present in the data, Data Management contractors consulted with the National Centre regarding these missing data. If National Centres were able to resolve these issues (e.g. participant response information was mistakenly miscoded or not entered into the DME software), information was provided to the data management contractor through the submission of an updated or revised DME database, and the necessary steps for preprocessing and preprocessing were completed. If the reported data inconsistencies were resolved, the scoring process was complete and the data proceeds to the next phase of processing.

Background Questionnaire processing

After all cognitive processing is completed, the data moved to the next phase– Background Questionnaire processing.

Harmonisation and National Background Variable Adaptations

As mentioned earlier in this chapter, although there was the essential need for standardisation across countries, countries did have the opportunity to modify or adapt background questionnaire variable stems and response categories to reflect national specificities. These modifications were referred to in general as 'national adaptations' of background questionnaire questions. As a result, changes to variables proposed by a National Centre occurred during the translation and adaptation process. Additionally, the Background Questionnaire (BQ) contractors agreed upon adaptations for questionnaire variables. These discussions regarding adaptations happened in the 'negotiation' phase between the country and the BQ contractor as well as the translation verification contractor prior to data collection. All changes and adaptations to questionnaire variables were captured in the Questionnaire Adaptation Sheet (QAS). It was the role of the BQ contractor to utilise the country's QAS file to approve national adaptations and create the harmonisation code for Data Management contractors.

For PISA-D processing, harmonisation or harmonising variables is a process of mapping the national response categories of a particular variable into the international response categories in order to be compared and analysed across countries. National Centres documented and implemented all required background variable adaptations in the following resources: QAS and the DME.

Any issues concerning national adaptations were handled by the country as well as by both the BQ contractor and the Data Management contractor. As the questionnaire experts, however, the BQ contractor provided the harmonisation code and approval for all changes to the harmonisation code for data processing. In order to verify the results of the harmonisation code, the BQ contractor was responsible for reviewing the harmonisation reports produced by ETS Data Management for any issues or concerns pertaining to national adaptations. In addition to the BQ contractor, the National Centres also reviewed these harmonisation reports and contacted both the BQ contractor and the Data Management contractor with approval or proposed changes. All variable change requests from the country were documented in the national harmonisation report.

Derivation coding

The derived variable code for the derivation of questionnaire variables was generated by the BQ contractor 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.

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 Centre. 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, proficiency level values, background questionnaire scales and weights.

The dynamic feature of the cleaning and verification software allowed for the Data Management contractor to tailor specific deliverables based the Consortium's data release timeline. In order to produce these customised files for the Consortium, each deliverable required a separate series of checks and reviews in order to ensure all data are handled appropriately and all values are populated as expected.

PREPARING FILES FOR PUBLIC USE AND ANALYSIS

In order to prepare for the public release of the PISA for Development Main Survey data, ETS Data Management provided data files in SPSS and SAS to National Centres 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 proficiency level values, weights and questionnaire scaled derived variables. During each of these phases of delivery, National Centres review 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 Main Survey, the following files were prepared:

- ***National Data File*** with all participant responses cognitive (raw, scored, timing) and background questionnaire items (responses and timing), PMK and screener data. These files included all raw variables, questionnaire scaled derived variables, sampling weights, replicate weights and proficiency level values.[2]
- ***International Database file*** a concatenated file of all countries to provide further information for analysis to National Centres. These data were separated into cognitive, questionnaire (including participant responses, PMK, proficiency levels and weights) and questionnaire timing data files.
- ***International Database file with variable suppressions*** a concatenated file produced after the first release of the International Database file. These data included all country requested variable suppressions. This data file was used for construction of the public use file (PUF) delivered to the OECD. These data were separated into cognitive, questionnaire (including respondent, PMK and household, proficiency levels and weights) and questionnaire timing data files.
- ***Analysis Reports*** were delivered by Data Management and Analysis and used by the Consortium and National Centres for quality control and validation. These reports highlighted performance results for groups, especially in the extent to which they agreed with expected outcomes. Additionally, the item analysis reports provided to countries included tables that contained summary information about the response types given by the respondents to the cognitive items. They also included, 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.

Data management and confidentiality, variable suppressions

For participation in PISA for Development, some country regulations and laws restrict the sharing of some data, as originally collected, with other countries. The goal of such disclosure control is to prevent the spontaneous or intentional identification of individuals in the release of data. On the other hand, suppression of information or reduction of detail clearly affects the analytical utility of the data. Therefore, both goals must be carefully balanced. As a general directive for

PISA data, the OECD requests that all countries make available the largest permissible set of information at the highest level of disaggregation possible.

According to the Technical Standards, each country provides the Consortium with early notification of any rules affecting the disclosure and sharing of PISA-D 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 change the response values must be applied prior to submitting data to the Consortium 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.

With the delivery of the data from the National Centre, the Data Management contractor 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. As part of PISA guidelines, country suppression requests generally involve specific variables that violate confidentiality and anonymity of participant data.[3]

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] Records excluded from the database include PISA-D Strand C participants that did not receive achievement levels or sampling weights due to specific disposition codes (i.e. refusal, incomplete, or other impairment).

[3] For the Strand C Main Survey, variables suppressed from the public release include Y016cQ01NA, Y016cQ11NA, Y23cQ07NA, Y23cQ08NA, Y027aQ09NA, HH009Q01NK and HH009Q01NL for Senegal.