10

Data Management Procedures

INTRODUCTION

The PISA assessment establishes standard data collection requirements that are common to all PISA participants. Test instruments include the same test items in all participating countries, and data collection procedures are applied in a common and consistent way amongst all participants to help ensure data quality. Test development is described in Chapter 2, and the data collection procedures are described in this chapter.

As well as the common test elements and data management procedures, the opportunity also exists for participants to adapt certain questions or procedures to suit local circumstances, and to add optional components that are unique to a particular national context. To accommodate the need for such national customisation, PISA procedures need to ensure that national adaptations are approved by the international contractor, are accurately recorded, and where necessary the mechanisms for re-coding data from national versions to a common international format are clearly established. The procedures for adapting the international test materials to national contexts are described in Chapter 2 and the procedures for adapting the questionnaires are described in Chapter 3. The mechanisms for re-coding data from national versions to a common international format are described in this chapter.

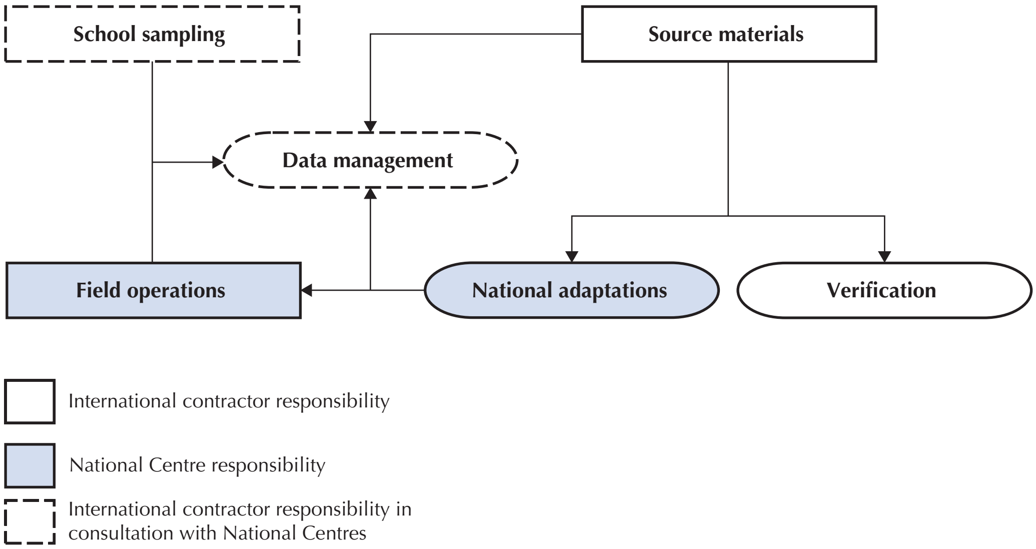
As well as planned variations in the data collected at the national level, the possibility exists for unplanned and unintended variations finding their way into the instruments. Data prepared by national data teams can be corrupted or inaccurate as a result of a number of unintended sources of error. PISA data management procedures are designed to minimise the likelihood of errors occurring, to identify instances where errors may have occurred, and to correct such errors wherever it is possible to do so before the data are finalised. The easiest way to deal with ambiguous or incorrect data would be to delete the whole data record containing values that may be incorrect. However, this should be avoided where possible since the deleted data records results in a decrease in the country's response rate. This chapter will therefore also describe those aspects of data management that are directed at identifying and correcting errors. These procedures applied for both the pencil and paper and computer-delivered components of PISA 2012.

The complex relationship between data management and other parts of the project such as development of source materials, instrument adaptation and verification, as well as school sampling are illustrated in Figure 10.1. Some of these functions are located within National Centres, some are located within the international contractor, and some are negotiated between the two.

Data management procedures must be shaped to suit the particular cognitive test instruments and background questionnaire instruments used in each participating country. Hence the source materials provided by the international contractor, the national adaptation of those instruments, and the international verification of national versions of all

■ Figure 10.1 ■

Data management in relation to other parts of PISA


School sampling
Source materials
Data management
Field operations
National adaptations
Verification
□ International contractor responsibility
▨ National Centre responsibility
⬚ International contractor responsibility in consultation with National Centres

instruments must all be reflected in the data management procedures. Data management procedures must also be informed by the outcomes of PISA sampling procedures. The procedures must reliably link data to the students from whom they came. Finally, the test operational procedures that are implemented by each National Centre, and in each test administration session, must be directly related to the data management procedures.

Figure 10.2 illustrates the sequence of major data management tasks in PISA, and shows the division of responsibilities between National Centres, the international contractor, and those tasks that involve negotiation between the two. This section briefly introduces each of the tasks. More details are provided in the following sections.

First, the international contractor provides the data management software *KeyQuest* to all National Centres. *KeyQuest* is a generic software that can be configured to meet a variety of data entry requirements. In addition to its generic features, the latest version of *KeyQuest* was pre-configured specifically for PISA 2012.

KeyQuest was preconfigured with all the PISA 2012 standard instruments: cognitive test booklets, background and contextual questionnaires, and student tracking instruments that are derived following implementation of the school sampling procedures. However, it also allows for instrument modifications such as addition of national questions, deletion of some questions and modification of some questions. A prerequisite for national modification of *KeyQuest* is international contractor approval of proposed national adaptations.

After the National Centres receive *KeyQuest*, they carry out student sampling and they implement *KeyQuest* modifications as a part of preparation for testing. By that time the variations from the core PISA sampling procedures such as national and international options (see Chapter 6) and the proposed national adaptations of the international source instruments (see Chapters 3 and 6) were agreed with the international contractor and all national versions of instruments have been verified.

Following test administration and coding of student responses, National Centres are required to enter the data into *KeyQuest*, to perform validity reports to verify data entry, and to submit the data to the international contractor.

As soon as data are submitted to the international contractor, additional checks are applied. During the process of data cleaning, the international contractor sends cleaning reports containing the results of the checking procedures to National Centres, and asks National Centres to clarify any inconsistencies in their database. In the questionnaires for example such inconsistencies might include the number of qualified teachers in a school exceeding the total number of teachers or unlikely (though not impossible) situations such as parents with higher degrees but no secondary education. The national data sets are then continuously updated according to the information provided by the National Centres. The cleaning reports are described in more detail below.

Once the international contractor has received all cleaning reports from the National Centres and has introduced into the database all corrections recommended in these reports, a number of general rules are applied to the small number of unresolved inconsistencies in the PISA database.

At the final data cleaning stage National Centres are sent the initial analysis reports containing cognitive test item information and frequency reports for the contextual questionnaires. The National Centres are required to review these reports and inform the international contractor of any inconsistencies remaining in the data. Further recoding is made after the requests from the National Centres are reviewed. At the same time sampling and tracking data is submitted and analysed and any consequential data recoding is implemented. At that stage the database is regarded as final, and ready for submission to the OECD.

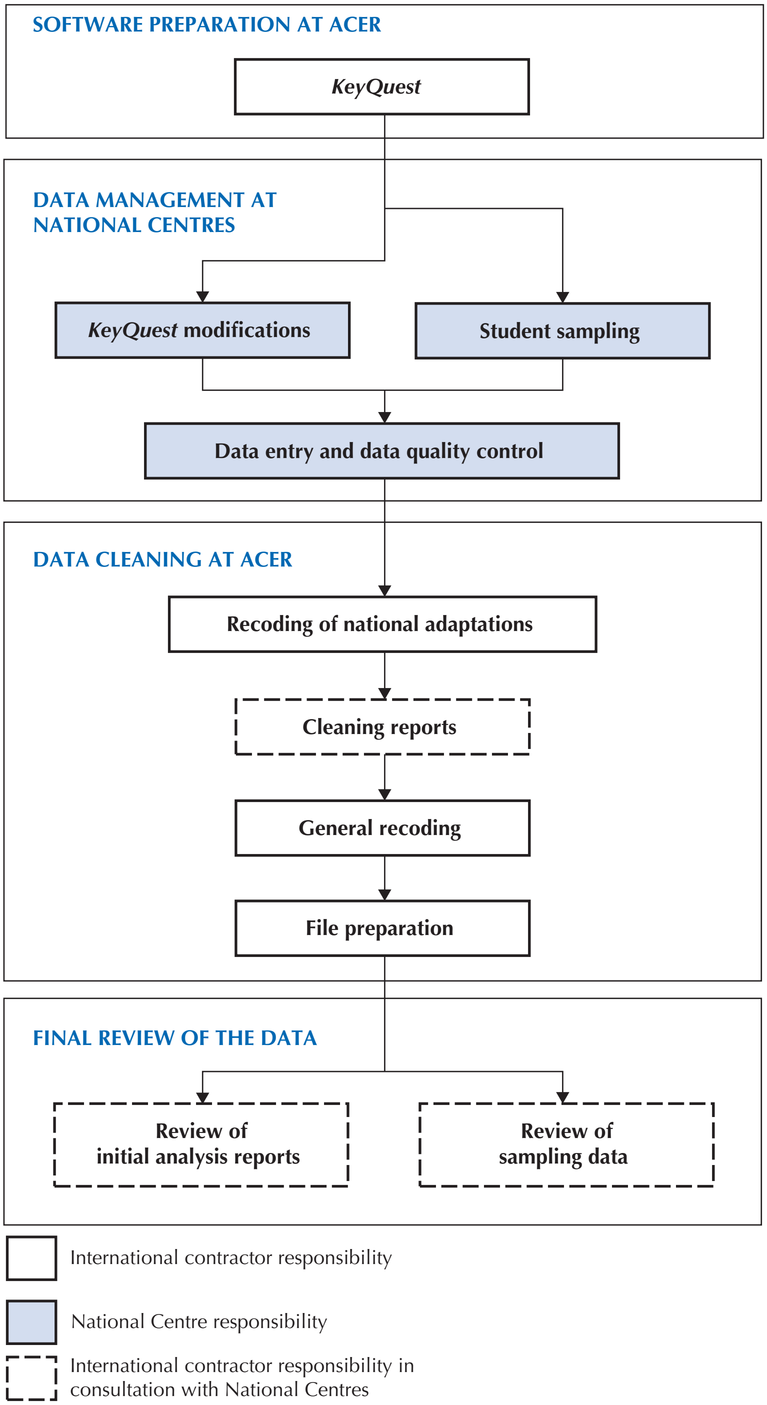
DATA MANAGEMENT AT THE NATIONAL CENTRE

National modifications to the database

PISA's aim is to generate comparable international data from all participating countries, based on a common set of test instruments. However, it is an international study that includes countries with widely differing educational systems and cultural particularities. Due to this diversity, some instrument adaptation is required. Hence verification by the international contractor of national adaptations is crucial (see Chapter 5). After adaptations to the international PISA instruments are agreed upon, the corresponding modifications in *KeyQuest* are made by National Centres.

■ Figure 10.2 ■

Major data management stages in PISA


SOFTWARE PREPARATION AT ACER
KeyQuest
DATA MANAGEMENT AT NATIONAL CENTRES
KeyQuest modifications
Student sampling
Data entry and data quality control
DATA CLEANING AT ACER
Recoding of national adaptations
Cleaning reports
General recoding
File preparation
FINAL REVIEW OF THE DATA
Review of initial analysis reports
Review of sampling data
International contractor responsibility
National Centre responsibility
International contractor responsibility in consultation with National Centres

Student sampling with *KeyQuest*

Parallel to the adaptation process, National Centres sample students using *KeyQuest*. The student sampling functionality of *KeyQuest* was especially developed for the PISA project. It uses a systematic sampling procedure by computing a sampling interval. *KeyQuest* samples students from the information in the list of schools. It automatically generates the *Student Tracking Form* and assigns one of the rotated forms of test booklets and questionnaires to each sampled student. For those countries that participated in the computer-based assessment option of PISA 2012, *KeyQuest* also samples

students from within the sample of students selected for the paper-based assessment and assigns one of the rotated computer-based forms to each sub-sampled student. For countries participating in the International Option of Financial Literacy (FL), the number of students to be sampled for paper-based assessment was increased in each sampled school so as to also achieve the required student sample size for FL. The extra students then assigned FL booklets. Thus, sets of students selected for FL and for main paper-based assessment do not overlap.

In the process of sampling, *KeyQuest* uses the study programme table, which defines the different study programmes operating in sampled schools and enables conversion of national variations into consistent international codes, and the sampling form designed for *KeyQuest,* which summarises all relevant school-level information required to facilitate selection of the school sample. These were agreed with the National Centres through a negotiation process managed through and recorded on the international contractor's website MyPISA (*http://mypisa.acer.edu.au*) and imported into *KeyQuest*.

Critical information about sampled students is captured in the Student Tracking Form in *KeyQuest*. It includes the information used in computing weights, exclusion rates, and participation rates. Other tracking instruments used in *KeyQuest* included the session report form which is used to identify the language of test for each student. The date of the testing session that the student attended is obtained from the session report and used in conjunction with the date of birth of the student entered on the tracking form to calculate the age of the student at the time of testing.

Data entry quality control

The national adaptation and student sampling tasks are performed by staff at each National Centre before testing. After testing, data entry and the running of *KeyQuest* validity reports are carried out by the National Centres.

Validation rules

During data entry, *KeyQuest* captures certain data entry errors through the use of validation rules that restrict the range and type of values that can be entered for certain fields. For example, for a standard multiple-choice item with four choices, one of the values of 1-4 each corresponding to one of the choices (A-D) that is circled by the student can be entered. In addition, code 9 (Missing) was used if none of the choices was circled and code 8 (Invalid) if two or more choices were circled. Finally code 7 (Not Applicable) was reserved for cases when a student was unable to provide a response through no fault of their own, such as when a poorly printed item presented to the student was illegible. Another example is a continuous variable for which it is reasonable to expect any answer from a student in the range from 0 to 200. The particular rule for this question would be to allow entry of values between 0 and 200 (inclusive), which is the range of valid responses for this item. It also allows entry of 9 999 which is used, in this case, to indicate missing data, 9 998 to indicate an invalid response and 9 997 to indicate that the question was not administered. The inbuilt validation rules ensured that no other codes could be entered.

Key violations

Furthermore, *KeyQuest* was programmed to prevent key violations. That is, *KeyQuest* was programmed to prevent the duplication of so called "keys", which are a combination of identifier codes. For example, a data record with the same combination of region, stratum and school identifiers could not be entered twice in the School Questionnaire instrument. A data record with a student ID that does not exist in Student Tracking Form could not be saved at data entry or imported into *KeyQuest* for user-assessable instrument. For more detailed information please refer to PISA 2012 *Data Management Manual.*[1]

KeyQuest also allows double entry of the test and questionnaire data and monitoring of the data entry operators. These procedures are described below.

Monitoring of the data entry operators

The data entry efficiency report was designed specifically for PISA 2012 to keep a count of all records entered by each data entry operator and the time required to enter them. The international contractor recommended for all countries to make use of these procedures for quality assurance purposes during data entry.

Double coding of occupational data

Another optional procedure for PISA 2012 was the double coding of occupational data. The double coding procedure allowed National Centres to check the validity of the data, as well as allowing identification of the areas where

supplementary coding tools could be improved. The main coding tool was the *International Standard Classification of Occupations: ISCO-08* (ILO, 2008)[2] with a small number of additional codes, described in the PISA 2012 *Data Management Manual*. The supplementary coding tools would typically include coding instructions, a coding index, and training materials developed at the National Centre.

Under this procedure, the occupational data from the student questionnaires and parent questionnaires (if applicable) were coded twice by different coders and entered into two *KeyQuest* tables specifically designed for this purpose. Following this the double entry discrepancies report was generated. The records for which there were differences between ISCO Codes entered into the two tables were printed on the report, analysed by the data manager and acted upon. The possible actions included making improvements to the coding instructions (if the same error was systematically produced by different coders), and/or providing further training for coders that were making more errors than others. Finally, the international contractor expected all discrepancies printed on the report to be resolved before the data were submitted.

The National Centres that participated in this option commented on the usefulness of the procedures for training of the coding staff. The possibilities for the international contractor to conduct a comprehensive analysis of these data were limited due to language constraints. However, one result that was observed was that those countries that required their coders to enter a word description as well as a four-digit code had fewer discrepancies than those that required only a four-digit code. This led to a reinforcement of the ILO (International Labour Organization) recommendation that procedures should involve entering occupation descriptions first and then coding them, rather than coding directly from the questionnaires.

Validity reports

After data entry was completed, National Centres were required to generate validity reports from *KeyQuest* and to resolve all discrepancies listed on these reports before submitting data to the international contractor.

The structure of the validity reports is illustrated by Figure 10.3. They include:

- comparison between tracking instruments and sampling verification (tracking instruments, sampling verification);
- data verification within tracking instruments (tracking instruments specific checks);
- comparison of the questionnaire and tracking data (Student Questionnaire - Student Tracking Form specific checks, identity checks questionnaires, identity checks occupation);
- comparison of the identification variables in the test data (identity checks booklets, identity checks CBA - computer-based assessment); and
- verification of the reliability data (reliability checks).

Some validity reports listed only incorrect records (e.g. students whose data were entered in more than one booklet instrument), whilst others listed both incorrect and suspicious records, which were records that could have been either correct or incorrect, but were deemed to be in need of confirmation. The resolution of discrepancies involved the following steps:

- correction of all incorrect records: e.g. students entered as ‘non participant’, ‘transferred out of school’ but who were also indicated on the Student Tracking Form as having been tested; and
- an explanation for the international contractor as to how records on the report that were listed as suspicious, but were actually correct, occurred (e.g. students with special education needs were not excluded because it is the policy of the school).

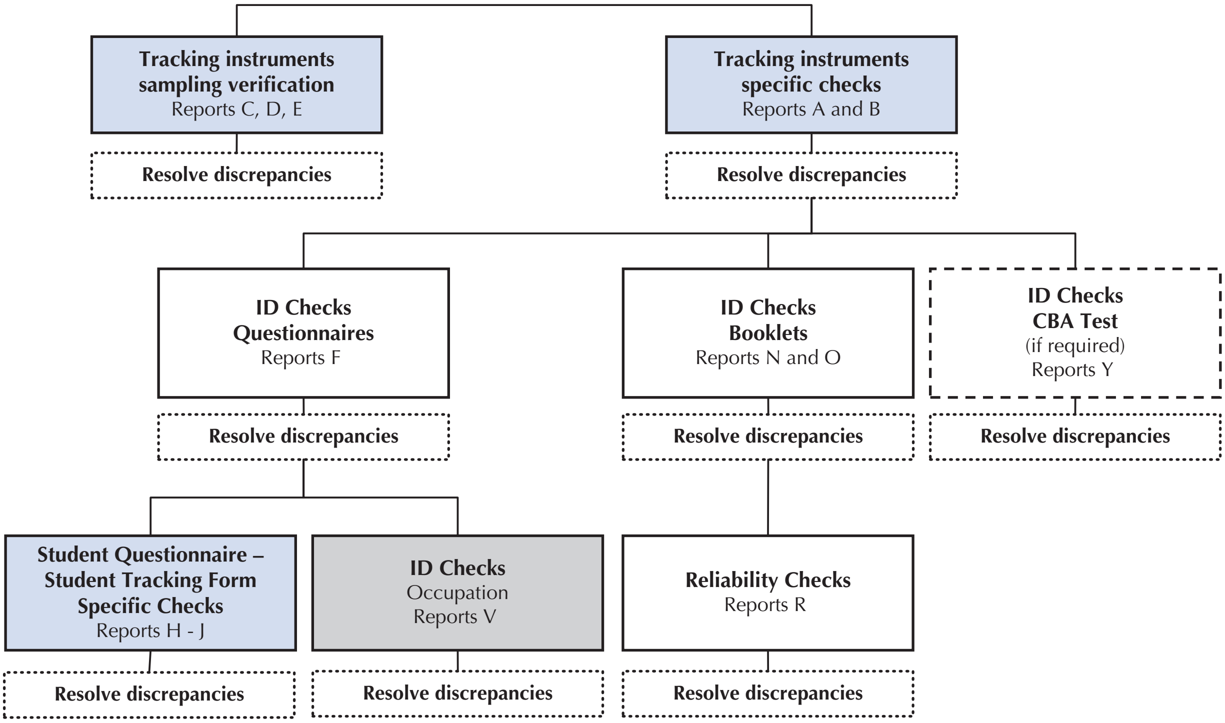
Due to the complexity and significant number of the validity reports, a validity report checklist was designed. More details about the validity reports can be found in the PISA 2012 *Data Management Manual*.

DATA CLEANING BY THE INTERNATIONAL CONTRACTOR

Recoding of national adaptations

When data were submitted by National Centres, the first step was to check the consistency of the database structure with the international database structure. An automated procedure was developed for this purpose. After data has been submitted, additional checks were undertaken. During that process of data cleaning, queries, if necessary, are sent to

■ Figure 10.3 ■
Validity Reports: General hierarchy



National Centres, along with reports containing data cleaning results. National Centres clarified all inconsistencies in their database. The national data sets were continuously updated and re-cleaned, to be consistent with information provided by the National Centres. For example, if a variable had been added to a questionnaire, the questionnaire adaptation sheet (which should have been agreed between the National Centre and international contractor before the modification was introduced into *KeyQuest*) was checked to find out whether this national variable required recoding into a corresponding international one, or had to be set aside as being for purely national use and returned to the country. Once all deviations were checked, necessary recodes for the submitted data to ensure they fit the international structure were performed. All additional or modified variables were set aside and returned to the National Centres in a separate file so that countries could use these data for their own purposes, but they were not included in the international database.

Data cleaning organisation

The data files submitted by National Centres often needed specific data cleaning or recoding procedures, or at least adaptation of standard data cleaning procedures. To reach the high quality requirements, the international contractor implemented dual independent processing; that is, two equivalent processing tools were developed – one in SPSS® and one in SAS® – and then used by two independent data cleaners for each dataset. The first step for one or both data cleaners was to check that discrepancies identified in Validity Reports had been resolved. Data cleaners checked all explanations provided in the Validity Reports and in the Item Information for Cleaning document (provided by National Centres at the data submission stage) and, if necessary, requested additional clarification from the relevant National Centres.

For each National Centre's data, two analysts independently cleaned all submitted data files, one analyst using the SAS® procedures, the other analyst using the SPSS® procedures. The results were compared at each data cleaning step for each National Centre. The cleaning step was considered complete for a National Centre if the recoded datasets were identical.

Computer-based assessment data

Countries that participated in the computer-based assessment option in PISA 2012 could elect to administer either the problem solving core component only or the problem solving core component plus the computer-based assessment of literacies (mathematics and reading) component. For countries that participated in the computer-based assessment option, the data files constructed from the test delivery and online coding systems were introduced into the data cleaning procedures. Student IDs from the data were checked against student IDs from the paper-based test data (from *KeyQuest*), although data from the computer-based assessment were retained even for those students who had not participated in the PISA paper-based test.

Cleaning reports

During the process of data cleaning, cleaning reports containing the results of the checking procedures were progressively sent to National Centres, with requests to clarify any inconsistencies in their database. The national data sets were then updated according to the information provided by the National Centre.

Many of the cleaning reports were designed to double check the *KeyQuest* validity reports run by the National Centres. If the data had been cleaned correctly at the National Centre, the cleaning reports would either not contain any records or would only list records that had been already explained in the *KeyQuest* validity reports. These cleaning reports were sent only to those countries whose data required additional cleaning.

However, there were additional checks that could not be conducted by the National Centres. For example, inconsistencies within the questionnaires could be checked only after the questionnaire data had been recoded back into the international format. Such cleaning reports were sent to all National Centres.

General re-coding

After receiving all cleaning reports from the National Centres and implementing the agreed corrections recommended in these reports, the international contractor applied the following general rules to the unresolved inconsistencies in the PISA database (this was usually a very small number of cases and/or variables per country, if any):

- Unresolved inconsistencies regarding student and school identification led to the deletion of the record in the database.
- The data of an unresolved systematic error for a particular cognitive item was replaced by the "Not Applicable" code. For instance, if a country informed the international contractor about a mistranslation or misprint for an item in the national version of a cognitive booklet, then the data for this item were recoded as "Not Applicable" and were not used in the subsequent analyses.
- If the country deleted a variable in the questionnaire, it was replaced by the "Not Applicable" code.
- If the country changed a variable in the questionnaire in such a way that it could not be recoded into the international format, the international variable was replaced by the "Not Applicable" code.

FINAL REVIEW OF THE DATA

As an outcome of the initial data cleaning, cognitive, questionnaire, and tracking data files were prepared for delivery to the OECD and for use in the subsequent analysis by National Centres and internationally.

Review of the test and questionnaire data

The final data cleaning stage of the test and questionnaire data was based on the review of national reports provided to each National Centre. After implementation of the corrections made in data cleaning reports and other general recoding, the international contractor sent initial analysis reports to every country, containing information about their test and questionnaire items, with an explanation of how to review these reports. For test items, the results of this initial analysis were summarised in six reports, four of which are described in Chapter 9. For the questionnaires, the reports contained descriptive statistics on every item in the questionnaire.

After review of those initial analysis reports, the National Project Manager (NPM) should have provided information to the international contractor about test items that appeared to have behaved in an unacceptable way (these are often referred to as "dodgy items") and any ambiguous data remaining in the questionnaires. Further recoding of ambiguous data followed. For example, if an ambiguity was due to printing errors or translation errors a "Not Applicable" code was applied to the item.

Recoding that was required following initial analyses of the international test and questionnaire data that were prepared for the OECD was introduced into the international data files.

Review of the sampling data

The final data cleaning step of the sampling and tracking data was based on analysis of the national tracking data files. The tracking files and sampling data for each country was checked and analysed and if required requested further recoding were implemented. For example, when a school was regarded as a non-participant because fewer than 25% of students from this school participated in the test, then all students from this school were deleted from the international database. Another example would be a school that was tested outside the permitted test window. All data for students from such a school would also be deleted.

NEXT STEPS IN PREPARING THE INTERNATIONAL DATABASE

When all data management procedures described in this chapter were complete, the database was ready for the next steps in preparing the public international database. Student weights and replicated weights were created as described in Chapter 8. Questionnaire indices were computed or scaled as described in Chapter 16. Cognitive item responses were scaled to obtain international item parameters that were used to draw plausible values as student ability estimates (see Chapters 9 and 12).

Notes

1. Available at *http://www.oecd.org/pisa*

2. In this section and throughout the document this edition is called ISCO-08 Manual for short. Available at : *http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm*

Reference

International Labour Organization (ILO) (2007), “ILO plans to support implementation of ISCO-08 in national and regional activities”, Paper for discussion by the United Nations Expert Group on International Economic and Social Classifications, New York, April 16-18, 2007.