Chapter 8

Creating and Checking the PIRLS International Database

Juliane Barth and Oliver Neuschmidt

8.1 Overview

The PIRLS 2006 International Database is a unique resource for policy makers and analysts, containing student reading achievement and background data from representative samples of fourth-grade students in 40 countries. Creating the PIRLS 2006 database and ensuring its integrity was a complex endeavor requiring close coordination and cooperation among the staff at the Data Processing and Research Center (DPC), the TIMSS & PIRLS International Study Center at Boston College, Statistics Canada, and the national centers of participating countries. The overriding concerns were to ensure that all information in the database conformed to the internationally defined data structure, that national adaptations to questionnaires were reflected appropriately in the codebooks and documentation, and that all variables used for international comparisons were indeed comparable across countries. Quality control measures were applied throughout the process to assure the quality and accuracy of the PIRLS data. This chapter describes the data entry and verification tasks undertaken by the National Research Coordinators (NRCs) and data managers of PIRLS participants, and the data checking and database creation procedures implemented by the IEA DPC in collaboration with the TIMSS & PIRLS International Study Center and Statistics Canada.

8.2 Software for Data File Creation

The IEA DPC went to great lengths to ensure that the data received from the PIRLS 2006 participants were of high quality and were internationally comparable. The foundation for quality assurance was laid before the first data arrived at the DPC by providing the PIRLS countries with software designed to standardize a range of operational and data related tasks.

- The WinW3S: Within-school Sampling Software for Windows (WinW3S) (IEA, 2005a) performed the within-school sampling operations adhering strictly to the sampling rules defined by the Statistics Canada and TIMSS & PIRLS International Study Center. The software also created all necessary tracking forms and stored student- and teacher-specific tracking form information (such as student's age, gender, and participation status).
- The WinDEM: Windows Data Entry Manager program (IEA, 2005b) enabled key entry of all PIRLS test and questionnaire data in a standard, internationally defined format. The software also includes a range of checks for data verification.

8.3 Data Entry at the National Centers

Each PIRLS 2006 national center was responsible for transcribing the information from the achievement booklets and questionnaires into computer data files. As described in Chapter 6, the IEA DPC supplied national research centers with the WinDEM software and manual (IEA, 2005b) to assist with data entry. The IEA DPC also provided countries with codebooks describing the structure of the data. The codebooks contained information about the variable names used for each variable in the survey instruments, and about field lengths, field locations, labels, valid ranges, default values, and missing codes. In order to facilitate data entry, the codebooks and data files were structured to match the test instruments and international version of the questionnaires. This meant that for each survey instrument there was a corresponding codebook, which served as a template for creating the corresponding survey instrument data file. The IEA DPC conducted a 3-day training seminar for the data managers from participating countries on the use of the WinW3S, WinDEM, and the codebooks.

The TIMSS & PIRLS International Study Center provided each NRC with the survey operations procedures, including general instructions about the

within-school sampling, translation and verification of test instruments, test administration, scoring procedures, and data entry and verification procedures (PIRLS 2005).

The national center in each country gathered data from tracking forms that were used to record information on students selected to participate in the study, as well as their schools, and teachers. Information from tracking forms was entered with help of WinW3S. The responses from the student achievement booklets as well as student, parents, teacher, and school questionnaires were entered into computer data files created from the codebook templates.

8.4 Data Checking and Editing at the National Centers

Before sending the data to the IEA DPC for further data processing, countries were responsible for checking the data files with the checks incorporated in WinDEM and specifically prepared for PIRLS 2006 and for undertaking corrections as necessary. The checks were mandatory for all countries:

- The structure of the data files conforms to the specifications in the international codebooks;
- The data values of categorical variables conform to the range validation criteria specified in the international codebooks;
- There are no duplicate records in the data file;
- There are no column shifts in the data file;
- The availability of the data is consistent with the corresponding indicator variables; and
- All participating schools, teachers, and students that have been selected are represented in the data files in accordance with the information in the survey tracking forms.

8.5 Submitting Data Files and Data Documentation to the IEA DPC

The following data files were used during data entry and submitted to the IEA DPC:

- The WinW3S database contained sampling information, as well as tracking form information (such as student's age, gender, and participation status), from all sampled students, teachers, and schools.

- The student background data file contained data from the *Student Questionnaire.*
- The parent (home) background data files contained data from the *Learning to Read Survey.*
- The student achievement data file contained student responses to the assigned test booklets.
- In order to check the reliability of the constructed-response item scoring, the constructed-response items were scored independently by a second scorer in a random sample of 100 booklets per type.[1] WinW3S defined the random sample. The responses from these booklets were stored in a reliability scoring file.
- The teacher background data files contained data from the *Teacher Questionnaire.*
- The school data file contained data from the *School Questionnaire*.

In addition to the submission of their survey data files to the IEA DPC, countries were requested to provide detailed data documentation. This included copies of all original survey tracking forms, copies of the national versions of translated test booklets and questionnaires, and National Adaptation Forms documenting all country-specific adaptations to the test instruments (for a description of authorized adaptations, see Chapter 5).

Countries also were asked to submit 100 test booklets of each type, which had been selected for the double scoring of constructed-response items. These booklets will be used to document the trend reliability of the scoring process between PIRLS 2006 and future cycles of the study.

8.6 Creating National Data Files for Within-country Analysis

Once the data were entered into data files at the national center, the data files were submitted to the IEA DPC for checking and input into the international database. This process is generally referred to as data cleaning. A study as complex as PIRLS required a complex data cleaning design. To ensure that programs ran in the correct sequence, that no special requirements were overlooked, and that the cleaning process ran independently of the persons in charge, the following steps were undertaken by the IEA DPC:

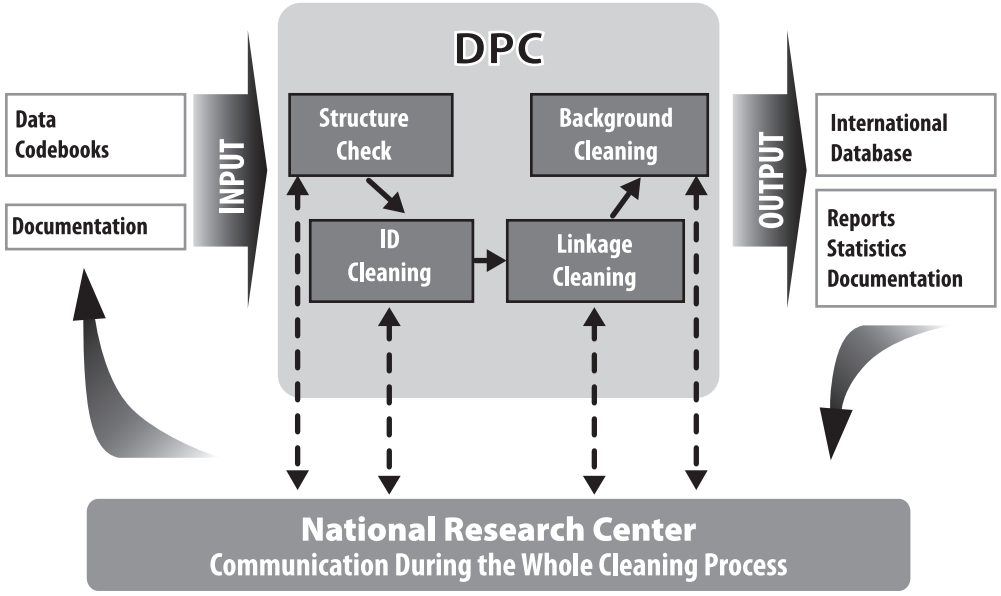
1 Booklet 9 and the Reader were exceptions, as they included only released texts from PIRLS 2006.

- Before use of real data, all data-cleaning programs were thoroughly tested using simulated data sets containing all possible problems and inconsistencies.
- All incoming data and documents were documented into a specific database. The date of arrival was recorded, along with any specific issues meriting attention.
- The cleaning was organized following strict rules. Deviations in the cleaning sequence were not possible, and the scope for involuntary changes to the cleaning procedures was minimal.
- All corrections to a country's data files were listed in a country-specific cleaning report.
- Occasionally, it was necessary to make changes to a country's data files. Every "manual" correction was logged using a specially developed editing program, which recorded all changes and allowed IEA DPC staff to undo changes, or to redo the whole manual cleaning process automatically at a later stage of the cleaning.
- Data Correction Software was developed at the IEA DPC and distributed among the participating countries to assist them in identifying and correcting inconsistencies between variables in the background questionnaire files.
- Once data cleaning was completed for a country, all cleaning steps were repeated from the beginning to detect any problems that might have been inadvertently introduced during the cleaning process.
- All national adaptations that countries recorded in their documentation were verified against the structure of the national data files. All deviations from the international data structure that were detected were recorded in a National Adaptation Database. The content of this database is available for data analysts as a supplement to the *PIRLS 2006 User Guide for the International Database* (Foy & Kennedy, 2008).

The main objective of the process of data checking and editing at the IEA DPC was to ensure that the data adheres to international formats, that school, teacher, parents and student information could be linked between different survey files, and that the data accurately and consistently reflected

the information collected within each country. Exhibit 8.1 presents a graphical representation of PIRLS data processing.

Exhibit 8.1 Overview of Data Processing at the IEA DPC



The program-based data cleaning consisted of the following steps:

- Documentation and structure check,
- Valid range check,
- Identification variable (ID) cleaning,
- Linkage check, and
- Resolving inconsistencies in background questionnaire data.

8.6.1 Documentation and Structure Check

For each country, data cleaning began with an exploration of its data file structures and a review of its data documentation: National Adaptations Forms, Student Tracking Forms, Teacher Tracking Forms, and Test Administration Forms. Most countries sent all required documentation along with their data, which greatly facilitated the data checking. The IEA DPC contacted those countries for which documentation was incomplete and obtained all forms necessary to complete the documentation.

The first checks implemented at the IEA DPC looked for differences between the international file structure and the national file structure. Some adaptations (such as adding national variables, or omitting or modifying international variables) were made to the background questionnaires in some countries. The extent and the nature of such changes differed across the countries. Some countries administered the questionnaires without any changes apart from translation, whereas other countries inserted items or options within existing international variables or added national variables. To keep track of any adaptations, National Adaptation Forms were used to adapt the codebooks, and, where necessary, the IEA DPC modified the structure of the country's data to ensure comparability with the structure of the international codebooks.

As part of this standardization process, since direct correspondence between the data entry instruments and the data files was no longer necessary, the file structure was rearranged from a booklet-oriented model designed to facilitate data entry to an item-oriented layout more suited to data analysis. Variables created purely for verification purposes during data entry were dropped at this time, and a provision was added for new variables necessary for analysis and reporting (i.e., reporting variables, derived variables, sampling weights, and achievement scores).

After each data file matched the international standard, as specified in the international codebooks, a series of standard cleaning rules were applied to the files. This was conducted using the set of programs developed at the IEA DPC that could identify and, in many cases, correct inconsistencies in the data. Each problem was recorded in a database, identified by a unique problem number, together with a description of the problem and the action taken.

Problems that could not be addressed were reported to the responsible NRC so that original data collection instruments and tracking forms could be checked to trace the source of the discrepancies. Wherever possible, staff at the IEA DPC suggested a remedy, and data files then were updated to reflect the solutions. After all automatic updates had been applied, remaining corrections to the data files were modified directly by keyboard, using a specially developed editing program.

8.6.2 Valid Range Check

"Valid range" indicates the range of the values considered to be correct and meaningful for a specific variable. For example, students' gender had two valid values: "1" for girls and "2" for boys. All other values were considered invalid. There also were questions in the school and teacher background questionnaires where the respondent wrote in a number—the principal was asked to supply the school enrollment, for example. For such variables, valid ranges may vary from country to country, and the broad ranges were set as acceptable to accommodate variations. It was possible for countries to adapt these ranges according to their needs, although countries were advised that a smaller range would decrease the possibility of mispunches. Data cleaning at the IEA DPC did not take smaller national ranges into account. Only if values were found outside the international accepted range were the cases mentioned in the list of inquiries sent to the countries.

8.6.3 Identification Variable (ID) Cleaning

Each record in a data file should have a unique identification number (ID). Duplicate ID numbers imply an error of some kind. If two records shared the same ID and contained exactly the same data, one of the records was deleted and the other remained in the data file. In the rare case that records contained different data apart from the ID numbers, and it was not possible to detect which records contained the "true data", both records were removed from the data files. However, the IEA DPC made every effort to keep such losses to a minimum.

The ID cleaning focused on the student background questionnaire file, because most of the critical variables were present in this file type. Apart from the unique ID, there were variables pertaining to students' participation and exclusion status, as well as dates of birth and dates of testing used to calculate age at the time of testing. The Student Tracking Forms were essential in resolving any anomalies, as was close cooperation with National Research Coordinators. The information about participation and exclusion was sent to Statistics Canada, where it was used to calculate participation rates, exclusion rates, and sampling weights.

8.6.4 Linkage Check

In PIRLS, data about students and their homes, schools, and teachers appear in several data files. It is crucial that the records from these files are linked to each other correctly to obtain meaningful results. Therefore, another important check

run at the IEA DPC is the check for linkage between the files. The students' entries in the achievement file and in the student background file must match one another, the home background file must match the student file, the reliability scoring file must represent a specific part of the student achievement file, the teachers must be linked to the correct students, and the schools must be linked to the correct teachers and students. The linkage is implemented through a hierarchical ID numbering system incorporating a school, class, and student component,[2] and is cross-checked against the tracking forms.

8.6.5 Resolving Inconsistencies in Background Questionnaires

All background questionnaire data were checked for consistency among the responses given. The number of inconsistent and implausible responses in background files varied from country to country, but considering the complexities involved, no country submitted data completely free of inconsistent responses. Inconsistencies were addressed on a question-by-question basis, using available documentation to make an informed decision. For example, question number 1 in the *School Questionnaire* asked for the total school enrollment (number of students) in all grades, while question 2 asked for the enrollment in the fourth grade only. Clearly, the number given should not exceed the number given for 1. All such inconsistencies that were detected were flagged, and the NRCs were asked to investigate. Those cases that could not be corrected and where the data made no sense were recoded to "Omitted".

Occasionally, filter questions with "Yes" or "No" answers were used to direct respondents to a particular section of the questionnaire. These filter questions and the following dependent questions were subjected to the following cleaning rule: If the answer to the filter question was "No" and yet the dependent questions were answered, then the filter question was recoded to "Yes". During data entry, dependent variables were not treated differently from others. However, a special missing code was applied ("Not applicable") to dependent variables during data processing.

Split variable checks were applied to questions where the answer was coded into several variables. For example, question 21 in the *Student Questionnaire* asked students to respond "Yes" or "No" to each item in a list of home possessions. Occasionally, students responded to the "Yes" boxes, but left the "No" boxes blank. Since in these cases it was clear that no response meant "No", these were recoded accordingly.

2 The ID of a higher level is repeated in the ID of a lower sampling level: The class ID holds the school ID, and the student ID contains the class ID (e.g., student 1220523 can described as student 23 in class 5 in school 122).

For further details about the standard cleaning procedures, please refer to the *General Cleaning Documentation PIRLS 2006* (IEA, 2007).

8.6.6 National Cleaning Documentation

National Research Coordinators received a detailed report of all problems identified in their data. This included documentation of any data problems detected by the cleaning programs and the steps applied to resolve them. NRCs also received a record of all deviations from the international data collection instruments and the international file structure

Additionally, the IEA DPC provided each NRC with revised data files incorporating all agreed upon edits, updates, and structural modifications. The revised files included a range of new variables that could be used for analytic purposes. For example, the student files included nationally standardized reading scores that could be used in preliminary national analyses to be conducted before the PIRLS 2006 International Database became available.

8.7 Handling of Missing Data

When the PIRLS data were entered using WinDEM, two types of entries were possible: valid data values and missing data values. Missing data can be assigned a value of omitted or not administered during data entry.

At the IEA DPC, additional missing codes were applied to the data to be used for further analyses. In the international database, four missing codes are used:

- Not administered: the respondent was not administered the actual item, and thus had no chance to read and answer the question (assigned both during data entry and data processing).
- Omitted: the respondent had a chance to answer the question, but did not do so (assigned both during data entry and data processing).
- Logically not applicable: the respondent answered a preceding filter question in a way that made the following dependent questions not applicable to him or her (assigned during data processing only).
- Not reached (only used in the achievement files): this code indicates those items not reached by the students due to a lack of time (assigned during data processing only).

8.8 Data Products

8.8.1 Data Almanacs and Item Statistics

Each country received a set of data almanacs, or summary statistics, produced by the TIMSS & PIRLS International Study Center. These contained weighted summary statistics for each participating country on each variable included in the survey instruments. These data almanacs were sent to the participating countries for review. When necessary, they were accompanied by specific questions about the data presented in them. They were also used by the TIMSS & PIRLS International Study Center during the data review and in the production of the reporting exhibits.

Each country also received a set of preliminary national item and reliability statistics for review purposes. The item statistics contained summary information about items characteristics, such as the classical item difficulty index, the classical item discrimination index, the Rasch item difficulty, and the Rasch mean square fit index. The reliability statistics contained summary statistics about the percent of agreement between scorers on the scores assigned to the item.

8.8.2 Versions of the National Data Files

Building the international database was an iterative process. The IEA DPC provided NRCs with revised versions of their country's data files whenever a major step in data processing was completed. This also guaranteed that the NRCs had a chance to review their data and run their own checks to validate the data files. Several versions of the data files were sent to each country before the PIRLS 2006 International Database was made available. Each country received its own data only. The first version was sent as soon as the data could be regarded as 'clean' concerning identification codes and linkage issues. These first files contained nationally standardized achievement scores calculated by the IEA DPC using a Rasch-based scaling method. Documentation, with a list of the cleaning checks and corrections made in the data, was included to enable the National Research Coordinator to review the cleaning process.

Updated versions of data almanacs were posted at regular intervals on the Internet by the TIMSS & PIRLS International Study Center for statistical review. A second version of the data files was sent to the NRCs when the weights and the international achievement scores were available and had been merged to the files. A third version was sent after all exhibits of the international report

had been verified and final updates to the data files had been implemented, to enable the NRCs to validate the results presented in the report.

8.9 The PIRLS 2006 International Database

The international database incorporates all national data files. Data processing at the IEA DPC ensured that:

- Information coded in each variable is internationally comparable;
- National adaptations are reflected appropriately in all variables;
- Questions that are not internationally comparable have been removed from the database;
- All entries in the database can be linked to the appropriate respondent—student, parents, teacher, or principal; and
- Sampling weights and student achievement scores are available for international comparisons.

In a joint effort of the IEA DPC and the TIMSS & PIRLS International Study Center at Boston College, a national adaptations database was constructed to document all adaptations to background questionnaires, including a description of how the adaptations were addressed in the international database, such as recoding requirements. The information contained in this database is provided in Supplement 2 of the *PIRLS 2006 User Guide for the International Database* (Foy & Kennedy, 2008). This accompanying documentation listing all national deviations from the international version of the background instruments will help analysts interpret the results correctly.

References

Foy, P., & Kennedy, A.M. (Eds.). (2008). *PIRLS 2006 user guide for the international database.* Chestnut Hill, MA: Boston College.

IEA. (2005a). WinW3S: Within-school sampling software for Windows. [Computer software and manual.] Hamburg: IEA Data Processing and Research Center.

IEA. (2005b). WinDEM: Data entry manager for Windows. [Computer software and manual.] Hamburg: IEA Data Processing and Research Center.

IEA. (2007). *General cleaning documentation, PIRLS 2006.* Hamburg: IEA Data Processing and Research Center.

TIMSS & PIRLS International Study Center. (2005). *PIRLS 2006 survey operations procedures.* Chestnut Hill, MA: Boston College.