8

Creating and Checking the PIRLS Database

Ursula Itzlinger

Knut Schwippert

8.1 Overview

Creating the PIRLS 2001 database, and ensuring its integrity, was a complex endeavor – requiring close coordination and cooperation among the staff at the IEA Data Processing Center (DPC), the PIRLS International Study Center at Boston College (ISC), Statistics Canada, and the national research centers of the 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 and data entry managers of participating countries, the data checking and database creation procedures implemented by the IEA Data Processing Center, and the steps taken at all institutions to confirm the integrity of the international database.

Database construction began with each national research center entering the data collected in the PIRLS 2001 survey into data files following the standard international format. Before sending the files to the IEA DPC, national center staff applied a system of checks to verify the structure of the data files. Checking and editing the national data sets was a matter of cooperation between the national centers, the ISC, Statistics Canada, and the DPC team.

The IEA DPC was responsible for checking the data files and applying standard cleaning rules to verify the accuracy and consistency of the data. Any queries were addressed to the national research centers, and modifications were made to the data files as necessary. The IEA DPC produced summary statistics for all variables in the background and achievement data for the national research centers, which were then reviewed by the ISC for any apparent oversights in recoding or valid range issues.

After all modifications had been applied, all data were processed and checked again. This process of editing the data, checking the reports, and implementing corrections was repeated as many times as necessary until all data were consistent and comparable within and between countries.

In preparation for creating the international database, the IEA DPC provided data almanacs containing international univariate statistics and item statistics to the national centers so that they could examine their data from an international perspective. This was one of the most important checks (in terms of international comparability of the data). While in a national context some statistics may seem plausible, it may become apparent in comparing data across countries that such interpretations lead to dubious results in an international context, despite accurate translation of the questionnaires. Any such instances were addressed, and the corresponding variables were either recoded or subject to removal from the international database.

The final tasks of database construction included achievement scores and sampling weights, distributing national data files and documentation to each of the participating countries, and creating the international database. National research centers received their processed national databases approximately six months after arrival at the DPC. At the same time, processed data files also were sent to Statistics Canada for the calculation of sampling weights (see Chapter 9) and to the ISC, where the achievement scores were computed (see Chapter 12).

8.2 Data Entry at the National Research Centers

To assist with data entry, the IEA DPC supplied the DataEntryManager (WinDEM) software and manual (IEA, 2001b), and held a training session on the use of the software. The International Study Center provided each national research center with a *Manual for Entering the PIRLS Data* (PIRLS, 2001a), which details prescribed procedures for data entry and verification. In addition, the *Survey Operations Manual* (PIRLS, 2001b) includes directions for submitting the data files to the IEA DPC.

The data manager at each PIRLS national research center gathered data from tracking forms used to record information on students selected to participate in the study, as well as about their schools, teachers, and parents. Together with the responses from the student achievement booklets and student, teacher, school, and parent question-

naires, the information from the tracking forms were entered into computer data files. Codebooks specifying the standardized format and layout of the data were provided as a supplement to the WinDEM software and the *Manual for Entering the PIRLS Data* (PIRLS, 2001a). While strongly encouraged to use the recommended WinDEM software, a few participating countries elected to use a different data entry system. However, they were required to conform to all specifications established in the international codebooks.

In order to facilitate data entry, the codebooks and data files were structured to match the tests and questionnaires. This meant that for each survey instrument there was a corresponding data file and codebook. Furthermore, countries administering the test booklets or questionnaires in more than one language had to carefully prepare for data entry. They needed to determine whether the different versions of the test booklets or questionnaire could be entered into one database, or if they required one database for each version.

8.3 Data Checking and Editing at the National Centers

Before sending the data to the DPC for further data processing, countries were responsible for checking data files with programs specifically prepared for PIRLS and for making corrections as necessary. The first step was the application of the checking programs that are a feature of the WinDEM program. These tools are intended mainly to identify invalid data, but also can check the consistency between some basic variables. An important feature of WinDEM is the ability to check for unique identification codes. These checks were obligatory for all countries.

In the application of the LinkPIRL program (IEA, 2001c), the identification variables (student, teacher, class, or school ID) were checked against one another both within and between all files. Examples of linkage errors include: schools that were reported as non-participating, but for which there was a questionnaire in the teacher file; or students listed in the achievement files for whom there was no corresponding identification number in the background files. NRCs were asked to recheck their records, and resolve the problems identified in the within-country cleaning process.

8.4 Submitting Data Files to the IEA Data Processing Center

Each country was responsible for submitting six data files to the IEA Data Processing Center: the student background questionnaire file, student achievement file, home background file, teacher background file, school background file, and the constructed-response scoring reliability file. Countries administering the 1991 Reading Literacy Study test booklets and questionnaires submitted a seventh file: the 10-year trend study file. (For details of these files, see section 6.11.)

In addition to the data files, countries were required to submit copies of all tracking forms, copies of their national versions of translated test booklets and questionnaires,

Data Management Forms documenting all national adaptations to the background questionnaires, and those booklets selected for the double scoring of constructed-response items.

8.5 DPC Quality Assurance Program

The IEA DPC has established a Quality Assurance Program to ensure that data is of high-quality, and that it is internationally comparable. Quality assurance was initiated before the first data arrived at the DPC through the provision of software to countries participating in PIRLS.

- The W3S software (IEA, 2001a) performs within school sampling and creates the required tracking forms.
- The WinDEM (IEA, 2001b) program performs data entry and data quality checks.
- The LinkPIRL program allows the NRCs to perform consistency checks between files.

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:

- All incoming data and documents were read into a specific database. The date of arrival was stored, along with any specific issues, with the person in charge of monitoring the characteristics of the data and documents.
- Thorough testing of all cleaning programs took place prior to their implementation by means of simulated data sets containing all possible problems and inconsistencies.
- 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.
- Regular reviews of the country-specific data processing were done by a quality-assurance work group.
- A validity check was implemented for all cleaning steps, once the cleaning for a specific country was done. A country's data were virtually treated as new incoming data, and was again subjected to the entire cleaning process. There could be no new findings; all findings at this stage had already been justified.

A comparison was made between the original data set and the final, clean data set. Any changes in the data set had to be documented in the country's cleaning documentation.

8.6 Data Checking and Editing at the IEA Data Processing Center

Once the data were entered into data files at the national research center, the data files were submitted to the IEA Data Processing Center for checking and input into the international database. This process is generally referred to as data cleaning. The program-based data cleaning consisted of the following steps:

- Documentation and structure check
- Identification number cleaning and linkage check
- Valid range check and cleaning of inconsistencies within and between background files
- Quality control cleaning.

Special issues addressed by the IEA DPC during the cleaning process included the handling of missing data, and cleaning of Trends in IEA's Reading Literacy Study 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: Data Management Forms, Student Tracking Forms, Class Sampling 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 DPC looked for differences between the international file structure and national file structures. 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 nature of such changes differed across the countries: some countries administered the questionnaires without any changes (apart from the translations), whereas other countries inserted items or options within existing international variables or added entirely new national variables. To keep track of any adaptations, NRCs were asked to complete Data Management Forms as they adapted the codebooks. Where necessary, the DPC modified the structure of the countries' data to ensure that the resulting data remained comparable between countries.

8.6.2 ID Cleaning and Linkage Check

Each record in a data file should have a unique identification number. 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 database. If the records contained different data apart from the ID, and it was impossible to detect which record contained the "true data," both records were removed from the database. The DPC tried to keep losses at a minimum, and, in only in a few cases, were data actually deleted.

The ID cleaning focused on the student background questionnaire file, because most of the critical variables were present in this file. Apart from the unique student ID, there were variables pertaining to the 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[1] were

1 Tracking Forms are used to record the sampling of schools, classes, teachers, and students. (see also Chapter 6).

essential in resolving any anomalies, as was close cooperation with NRCs (in most cases, the Student Tracking Forms were completed in the country's official language). The information about participation and exclusion was sent to Statistics Canada, where it was used to calculate students' participation rates, exclusion rates, and student sampling weights.

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

8.6.3 Valid Range Check, Filter-Dependent Check, and Consistency Check

"Valid range" indicates the range of values considered to be correct and meaningful for a specific variable. For example, the student gender variable had two valid values: "1" for a girl, and "2" for a boy. All other values are invalid. There were also questions in the school and teacher questionnaires for the respondent to write in a number – for example, the principal was asked to supply the school enrollment. For such variables, valid ranges may vary from country to country, and the acceptable ranges were set very wide 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. Cleaning at the 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 countries. In cases where out-of-range values were found in the achievement file, the data were set to "Omitted" if the true value could not be retrieved.

Filter questions, which appear in some questionnaires, were used to direct the respondent to a particular section of the questionnaire. Depending on the response to a filter question, responses to subsequent questions are either expected or not expected. During data entry, these dependent

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 be described as student 23 of class 5 in school 122).

variables are not treated differently from any others. However, a special missing code is applied to dependent variables during data processing (for details on the handling of missing data, see section 8.6.5).

The number of inconsistent and implausible responses in background files varied from country to country, but no country's data was completely free of inconsistent responses. Treatment of these responses was determined on a question-by-question basis, using available documentation to make an informed decision. One example of inconsistencies between files is when a school principal states that his or her school has no library, but the teacher in the same school indicates that students are taken to the school library regularly. These cases were not changed in either file, provided mispunches were ruled out as cause.

8.6.4 Quality Control Cleaning

Quality control cleaning ensures that all necessary recoding of variables was performed correctly, and that consistency within and between files could be verified. The variables in the database have complex interrelationships. To avoid changes that make the relationship between two variables consistent but breaks the relationship with a third variable, a final cleaning step was established to take care of such multiple relationships within the database. This quality control cleaning can be interpreted as a check of the results of all earlier checks. After this variable-level cleaning, the consistency check between files was performed.

8.6.5 Handling of Missing Data

When the PIRLS data were entered using WinDEM, two types of entries were possible: valid data values or missing data values. Missing data can be assigned a value of omitted, not administered, or invalid 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, five missing codes are used:

- Not administered – the respondent was not administered the actual item. He or she 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).

- Not interpretable (only used in the achievement files) – this code was used for multiple-choice items that were answered, but the chosen answer options were not clear – as well as for constructed-response items where the scorer assigned two or more scores (assigned during data entry and data processing).

8.6.6 Specific Cleaning Issues of the Trends in IEA's Reading Literacy Study

The Trends in IEA's Reading Literacy Study is a repetition of the IEA's 1991 Reading Literacy Study. Nine of the countries that participated in the 1991 study elected to re-administer the test in 2001 (for a list of these countries, see Exhibit 5.4). The requirements for the Trends in IEA's Reading Literacy Study were that the achievement test and the student background questionnaires must be administered in exactly the same way, and that the cleaning procedures be applied in the same way as in 1991.

As a result, data cleaning for the Trends in IEA's Reading Literacy Study data is somewhat different in comparison to the cleaning rules for PIRLS (International Association for the Evaluation of Educational Achievement, 1995):

- All items following the last item containing a valid value were recoded to "Not reached."
- An additional missing value, "Invalid," indicates that the data were recorded in an invalid or inconsistent way. This value was used only in the student background file. A more detailed description of the Trends in IEA's Reading Literacy Study data cleaning can be found in the cleaning documentation of PIRLS 2001 (Barth, Itzlinger, Niemeyer, & Schwippert, 2001).

8.7 Returning Data to National Centers

As soon as the ID cleaning was complete, and the file structures had been standardized, participating countries received their national data files back from the DPC, in order to conduct preliminary national analyses. These preliminary data sets did not include national variables, derived variables, scaled scores, or sampling weights. Due to the timelines in PIRLS, several versions of the data were sent to the national research centers, with each subsequent version containing more features.

When data processing was complete, final national data sets were sent to countries along with final sampling weights, international scores, derived variables, and all international and national variables. National variables were placed in extra files that could be merged with the files containing the international variables.

8.8 Creating the International Database

The international database incorporates all national data files. After data processing by the DPC, it can be 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, teacher, parent, or principal.
- Sampling weights and student achievement scores are available for international comparisons.

In a joint effort between the IEA DPC and the ISC at Boston College, a National Adaptations Database containing all adaptations to questionnaires made by individual countries (documenting how they were handled) was constructed. The meaning of country-specific items can also be found in this database, as well as recoding requirements of the ISC. Information contained in this database is provided in the user guide for the international database upon release of the PIRLS 2001 data.

The PIRLS 2001 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 from 35 countries. In all, the database contains more than 713 variables, with data from 5,777 schools, 7,041 teachers, 153,340 students, and 131,047 parents.

References

Barth, J., Itzlinger, U., Niemeyer, A. & Schwippert, K. (2001). *Cleaning documentation for PIRLS 2001*. Hamburg: Unpublished document, IEA Data Processing Center.

IEA. (1995). *The IEA Reading Literacy Study: technical report*. The Hague: IEA.

IEA. (2001a). *W3S: Within-school sampling software*. Hamburg: IEA Data Processing Center.

IEA. (2001b). *WinDEM: Software for data entry and verification*. Hamburg: IEA Data Processing Center.

IEA. (2001c). *LinkPIRL guide*. Hamburg: IEA Data Processing Center.

Progress in International Reading Literacy Study (PIRLS). (2001a). *Manual for entering the PIRLS data* (PIRLS Ref. No. 01-0004) Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001b). *Survey operations manual* (PIRLS Ref. No. 01-0001). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.