8

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.

naires, the information from the tracking
f

- 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

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 f3oon as th wnlg Data th hwditdth hwd/-
Reading77ing can be fD(e)251 -1.tir,l,sJT*(Niem,[s995):)]Trn (v)3

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

Bars