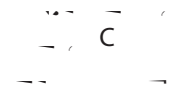# CHAPTER 9

Creating the PIRLS 2016 International Database

This chapter describes the procedures implemented by IEA Hamburg for verifying the PIRLS 2016 data and creating the PIRLS 2016 International Database (IDB).

Preparing the PIRLS 2016 International Database (IDB) and ensuring its integrity was a complex endeavor requiring extensive collaboration between IEA Hamburg, the TIMSS & PIRLS International Study Center, Statistics Canada, and the national centers of participating countries. Once the countries had created their data files and submitted them to IEA Hamburg, an exhaustive process of verification and editing known as "data cleaning" began.

Data cleaning is the process of checking data for inconsistencies and formatting the data to create a standardized output. The overriding concerns of the data cleaning process were to ensure:

- All information in the database conformed to the internationally defined data structure

- The content of all codebooks and documentation appropriately reflected national adaptations to questionnaires

- All variables used for international comparisons were in fact comparable across countries (after harmonization, where necessary)

- All institutions involved in this process applied quality control measures throughout in order to assure the quality and accuracy of the PIRLS 2016 data

IEA Hamburg was responsible for checking the data files from each country, applying standardized data cleaning rules to verify the accuracy and consistency of the data and documenting any deviations from the international file structure. Data files were created at each country's

to the filter question. For example, Question 9 of the school questionnaire reads: "Does your school have a school library? If yes, go to #9a, and if No, go to #10." If a respondent chooses "No," the online survey skips directly to Question 10, omitting Questions 9a and 9b. Not only does the skip logic save the respondents' time, it also results in fewer inconsistencies in the data received by IEA Hamburg.

Data Verification at the National Centers

Before sending the data to IEA Hamburg for further processing, national centers carried out mandatory validation and verification steps on all entered data and undertook corrections as necessary.

While the questionnaire data were being entered, the data manager or other staff at each national center used the information from the Tracking Forms to verify the completeness of the materials. Student participation information (e.g., whether a student participated in the assessment or was absent) was entered or imported into WinW3S.

The validation process was supported by an option in WinW3S to generate an inconsistency report. This report listed all of the types of discrepancies between variables recorded during the within-school sampling and test administration processes and made it possible to cross-check

Exhibit 9.1: Overview of Data Processing at IEA Hamburg
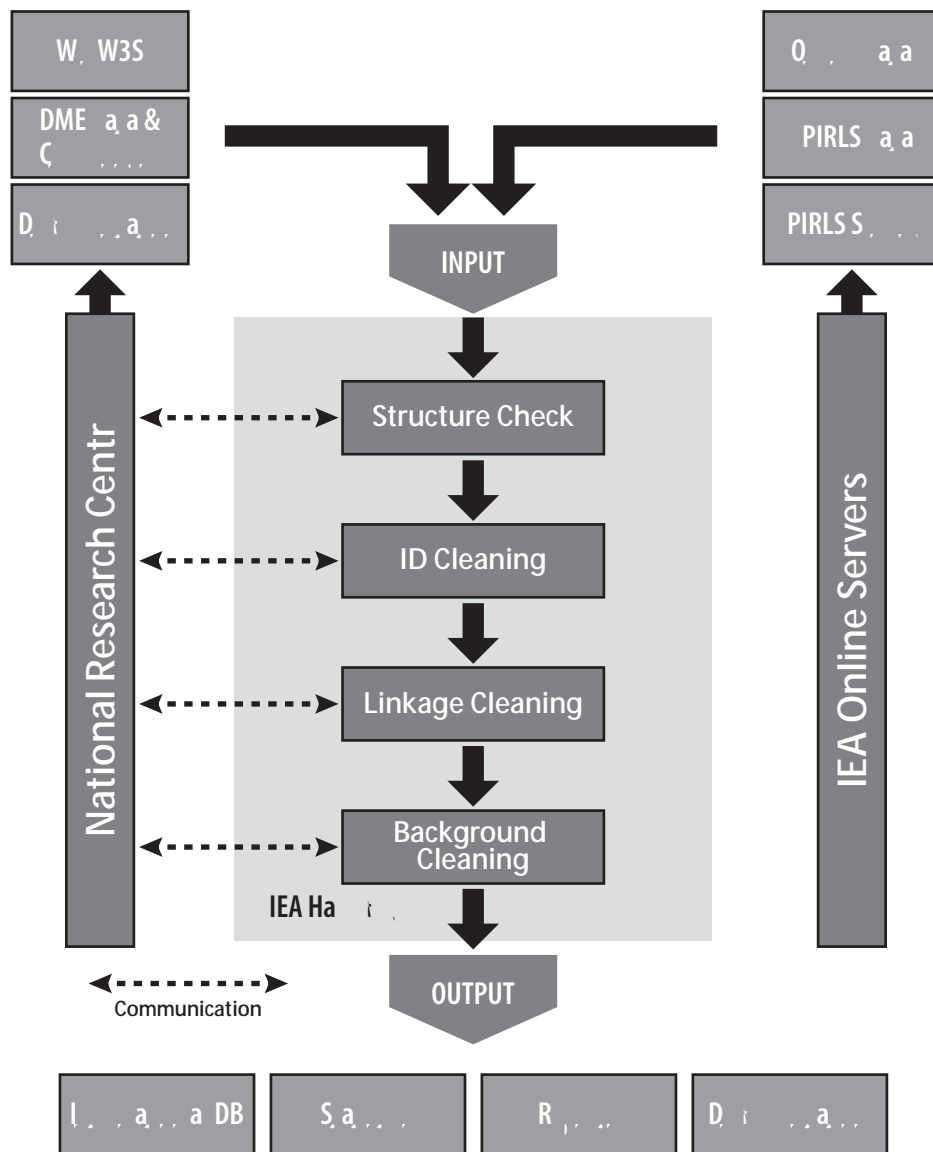
Structural Check

For each country, data cleaning began with a review of data file structures and data documentation, including a review of national adaptation comments, Student Listing Forms, Student Tracking Forms, Teacher Tracking Forms, and the Survey Activities Questionnaire.

After the review, IEA Hamburg merged the tracking information and sampling information captured in the WinW3S database with the student-level databases containing the corresponding student instrument data, and, if applicable, ePIRLS data. For countries administering questionnaires through online and paper modes, IEA Hamburg merged the questionnaire data across modes of administration. At this stage, data from the different sources were transformed and imported into one SQL database so that this information would be available during all further data processing stages.

Background Cleaning

The amount of inconsistent and implausible responses in questionnaire data files varied considerably across countries. IEA Hamburg determined the treatment of inconsistent responses on a question-by-question basis, using all available documentation to make an informed decision. IEA Hamburg staff also checked all questionnaire data for consistency across the responses given. For example, Question 1 in the school questionnaire asked for the total school enrollment in all grades, and Question 2 asked for the enrollment in the target grade only. Logically, the number given as a response to Question 2 could not exceed the number provided by school principals in Question 1. Similarly, it is not possible that the amount of 3e2a6 a renae ar 3(a)19s n33ss

Handling of Missing Data

Two types of entries were possible during the PIRLS 2016 data capture: valid data values and missing data values. Missing data can be assigned a value of "omitted/invalid" or "not administered"

weights and international achievement scores were available and had been merged with the data files. This version, sent to countries in advance of the 8th NRC meeting in Riga, Latvia in June 2017, contained only records that satisfied the sampling standards, and allowed the NRCs to replicate the results presented in the international reports.

Interim data products were accompanied by detailed data processing and national adaptation documentation, codebooks, and summary statistics. The summary statistics, preliminary versions of the PIRLS 2016 Almanacs, were created by the TIMSS & PIRLS International Study Center and included weighted univariate statistics for all questionnaire variables for each country. For categorical variables, representing the majority of variables, the percentages of respondents choosing each of the response options were displayed. For continuous numeric variables, various descriptive statistics were reported, including the minimum, maximum, mean, standard deviation, median, mode, and percentiles. For both types of variables, the percentages of missing data were reported. Additionally, for the achievement items, the TIMSS & PIRLS International Study Center provided item analysis and reliability statistics listing information regarding the number of valid cases, percentages, percent correct, Rasch item difficulty, scoring reliability, and so forth. These statistics were used for a more in-depth review of the data at the international and national levels in terms of plausibility, unexpected response patterns, etc. More information on reviewing item statistics is available in Chapter 10.

Final Product – the PIRLS 2016 International Databases

The data cleaning effort implemented at IEA Hamburg ensured that the PIRLS 2016 international databases contained high-quality data. More specifically, the process ensured that:

- Information coded in each variable was internationally comparable
- National adaptations were reflected appropriately in all variables
- All entries in the database could be successfully linked within and across levels
- Sampling weights and student achievement scores were available for international comparisons

Supplements to the *PIRLS 2016 User Guide for the International Database* document all national adaptations made to questionnaires by individual countries and how they were handled in the data. The meaning of country-specific items also can be found in this supplement, as well as recoding requirements by the TIMSS & PIRLS International Study Center.