Creating and Checking the TIMSS and PIRLS 2011 Databases

Oliver Neuschmidt

Preparing the international databases for the 2011 TIMSS and PIRLS assessments was a complex endeavor requiring extensive data checking and database creation procedures implemented by the IEA Data Processing and Research Center (DPC), the TIMSS & PIRLS International Study Center, Statistics Canada, and the national centers of participating countries. Once the countries had created their data files (see **Description of the TIMSS and PIRLS 2011 Data Files**) and submitted them to the IEA DPC, an exhaustive process of checking 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 that the data accurately reflect the information collected in each of the participating countries, that all information conformed to the international format, and that information from students, parents, teachers, and principals could be matched across different data files. This chapter describes the data cleaning procedures involved in the creation of the PIRLS and TIMSS 2011 database.

The data cleaning process required close collaboration between the national centers, the IEA Data Processing and Research Center (IEA DPC), the TIMSS & PIRLS International Study Center and Statistics Canada. IEA Data Processing and Research Center was the central hub of the data cleaning process. The IEA DPC 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 national center, and reviewed prior to submission to the IEA DPC. The National Research Coordinators (NRCs) collaborated with the IEA DPC to resolve any queries which emerged during the data cleaning process, and the NRCs checked interim versions of the national/benchmarking participant database(s) produced by the IEA DPC. The TIMSS & PIRLS International Study Center provided the NRCs with univariate data almanacs containing summary statistics on each variable so that the national centers could evaluate their data from an international perspective.

The TIMSS & PIRLS International Study Center also scaled the achievement and background data (please refer to **Scaling the TIMSS and PIRLS 2011 Achievement Data** and **Creating and Interpreting the TIMSS and PIRLS 2011 Context Questionnaire Scales**) and produced achievement scores (plausible values) and scores on the background scales. Using the tracking forms and data provided by the IEA DPC, Statistics Canada calculated the sampling weights, population coverage, and school and student participation rates, in collaboration with the IEA DPC.

Overview of the Data Cleaning Process

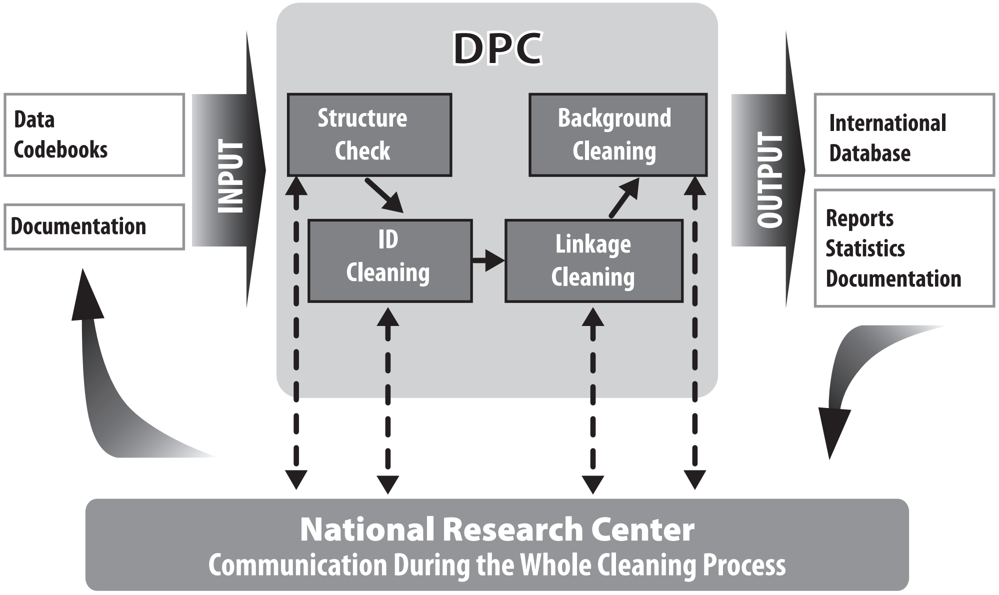
As illustrated in the following diagram, a uniform data cleaning process was followed, involving regular consultation between the IEA Data Processing and Research Center and the NRCs. Following data collection and data entry, each country submitted its data, codebooks and documentation to the IEA DPC. The IEA DPC, in collaboration with the NRCs, performed four procedures upon the submitted data and documentation:

(1) a structural check;
(2) a check of the identification (ID) variables;
(3) linkage cleaning;
(4) and background cleaning.

The cleaning process was an iterative process. Numerous cycles of the four-step cleaning procedure were completed on each national data set. The repetition ensured that all data were properly cleaned and that any new errors that could have been introduced during the data cleaning were rectified. The cleaning process was repeated as many times as necessary until all data were made consistent and comparable. Any inconsistencies detected during the cleaning process were solved in collaboration with national centers. All corrections made during the cleaning process were documented in a cleaning report, produced for each country.

After the final cleaning iteration, each country's data were sent to Statistics Canada for the calculation of sampling weights, and then the data, including sampling weights, were sent to the TIMSS & PIRLS International Study Center so that the scaling could be performed. The NRCs were provided with interim data products to review at three different points in the process.

Overview of data processing at the IEA Data Processing and Research Center



Creating the Data Files at the National Centers

Overall the national centers submitted data files of high quality and consistency to the IEA Data Processing and Research Center. As outlined in **Operations and Quality Assurance**, the data were first entered and then reviewed at the national centers before submission to the IEA DPC. Throughout the process, the NRCs were supported by **Operational Procedures manuals**, training sessions, and specialized software. In addition to facilitating the sampling process, the **WinW3S software** created by the IEA DPC assisted the NRCs in creating a consistent **hierarchical identification system** and providing all necessary tracking forms to store school, teacher and student-level information. **The WinDEM software**, also created by the IEA DPC, ensured internationally standardized data entry.

Using WinDEM, staff at each country's national center entered the data into **data files**. National centers were required to follow a stringent quality control process to ensure reliability of the data entry. Following the data entry, national centers were responsible for verifying their data sets and correcting inconsistencies (such as mismatched data across survey files, duplicate identification codes, out-of-range values, etc.). The WinDEM software included a range of data verification checks that facilitated the identification and correction of inconsistencies before data submission to the IEA DPC.

Upon completion of a thorough check of the data at the local level, the NRCs submitted the data set(s), codebooks and the accompanying documents to the IEA DPC. The documents submitted included National Adaptation Forms, Student Tracking Forms, and Student-Teacher Linkage Forms (TIMSS) or Student Listing Forms (PIRLS). Most countries submitted all required documentation along with their data, which greatly expedited the data checking process. Countries returning incomplete data or documentation were prompted to submit the outstanding material.

Receipt of the Data at the IEA Data Processing and Research Center

Upon completion of the data entry and quality control process, each country submitted its data and accompanying documentation to the IEA Data Processing and Research Center for data cleaning and formatting. Upon arrival at the IEA DPC, the data and documentation were registered in a special database developed to document the receipt of files. Each entry in this database included the date of arrival as well as any specific details related to the data cleaning. A second database was set up to record national adaptations to the data files. All adaptations to the international file structure were recorded in this database, including instructions for recoding the adapted data to match the international format.

In order to efficiently and accurately complete the data cleaning process, the IEA DPC developed an extensive suite of bespoke data processing programs using the Statistical Analysis Software (SAS) package. All data cleaning programs were thoroughly tested using simulated data sets before they were applied to the 2011 data.

Documentation and Structure Check

Data cleaning began with an inspection of the structure of the data files and a review of the documentation submitted by the NRCs. The documentation that was reviewed at this stage included the National Adaptation Forms, Student Tracking Forms, Student-Teacher Linkage Forms (TIMSS) or Student Listing Forms (PIRLS), and Test Administration Forms. The purpose of this review was to verify that all documentation had been submitted and properly completed. At the same time, the tracking and sampling information captured by the WinW3S database was merged with the achievement and questionnaire data stored in the WinDEM data files. The data from the school and teacher questionnaires were merged with the other files at this time. The merged data were then converted to SAS format in preparation for the data processing.

The first structure checks included a review of any differences between the national and international file structures. Although most countries administered the questionnaires without any major modifications, a number of countries made adaptations to the international instruments to incorporate national options for data collection. These could involve adding new national variables or inserting items or options within existing international variables, or omitting international variables from their questionnaires. NRCs making adaptations to the international instruments were required to follow strict guidelines specified in the Survey Operation Procedures and document the adaptations in the National Adaptation Forms.

All national adaptations recorded on the National Adaptation Forms were checked against the structure of the national data files. When possible, national adaptations were recoded to follow the international data structure. However, if international comparability could not be assured, the corresponding data was removed from the international database. All deviations from the international data structure were documented in the ***PIRLS 2011 International Database* and *User Guide*** (Foy & Drucker, 2013) and ***TIMSS 2011 International Database* and *User Guide*** (Foy, Arora, & Stanco, 2013).

Following the recoding of the national adaptations, the file structure for the achievement data was rearranged from a booklet-oriented format to an item-oriented format in order to make the data files more suitable for data management and international data analysis. In addition, variables created solely for verification purposes during the data entry stage were deleted.

When it had been established that each data file matched the international format, as specified in the international codebooks, a series of standard cleaning rules and consistency checks were applied, such as checking for out-of-range values. This procedure was conducted using the SAS cleaning program. Each problem encountered was recorded in a database, and identified by a unique problem number. The record in the database included a description of the problem and the action taken by the program to resolve it. All problems that could not be rectified automatically were reported to the NRC, and the original data collection instruments and tracking forms were checked to trace the source of the errors. The NRC was invited to propose to resolve the issues, although staff at the IEA DPC proposed a solution wherever possible. Often, the solution could be found within the data or the tracking forms. In the rare case that the errors could not be resolved, the IEA DPC applied a general cleaning rule to rectify the problem. After all issues had been resolved, IEA DPC staff used the bespoke SAS recoding scripts to apply any remaining corrections to the data files.

Identification (ID) Cleaning

Upon completion of the structural modifications, the identification (ID) variables in each data file were examined. Each record in the data file had a unique identification variable to identify, track and document each respondent. If two records shared the same ID number and contained exactly the same data, the IEA DPC deleted one of the records in the database (retaining only one of the two records). If both records had the same ID yet different data, and it was not possible for the IEA DPC and NRCs to identify which record contained the "true data," both records were removed from the database. The deletion of both records was exceptional, and in this cycle only a few cases were actually deleted.

Although the ID cleaning included all files, the ID cleaning effort focused on the student background questionnaire files, which contained most of the critical ID variables. Apart from the unique student ID numbers, it was necessary to check variables in these files pertaining to student participation and exclusion status.

Linkage Checks

As data on students, parents, teachers and schools appeared in a number of different data files, a process of linkage cleaning was implemented to ensure that the data files would correctly link together. The linking of the data files followed a hierarchical system of identification codes that included school, class, and student components (see **Linking Students to their Teachers and Classes**). These codes linked the students with their class and/or school membership. Linkage cleaning consisted of a number of checks to verify that student entries matched between achievement files, student background files, scoring reliability files and home background files (PIRLS countries). In addition, at this stage, checks were conducted to ensure that teacher and student records linked correctly with their corresponding schools. The Student Tracking Forms, Teacher Tracking Forms, and Student-Teacher Linkage Forms (TIMSS only) were crucial in resolving any anomalies. The IEA DPC also liaised with NRCs about any problematic cases, and the national centers were provided with standardized reports listing all inconsistencies identified within the data.

Resolving Inconsistencies in Background Questionnaire Data

After each file matched the international standard specified in the codebooks, and the identification and linkage cleaning was complete, a series of standard cleaning rules were applied to the background questionnaire data files. The cleaning program identified, and in many cases automatically corrected, inconsistencies in the data. When inconsistencies could not be reconciled automatically, they were resolved on a question-by-question basis using available documentation to make an informed decision. Among the inconsistencies encountered in the data were irregularities with regard to filter questions, imputation of implied answers by the respondents, data entry errors by the national staff, and other inconsistent response patterns (See **Resolving Inconsistencies in the TIMSS and PIRLS 2011 Data**).

The number of inconsistent or implausible responses in the data files varied from country-to-country, and each country and/or benchmarking participant had peculiarities which needed to be resolved. Each issue was recorded in a database. Issues that could not be solved using systematic cleaning rules were reported to the NRC so that the original questionnaires could be cross-checked. Where the national centers could not solve the problems by inspecting the instruments and forms, a set of final cleaning rules were applied.

Interim Data Products

Before the TIMSS and PIRLS International Databases were finalized, three major interim versions of the data files were sent to each country. Each country received its own data only. The first version was sent as soon as the data could be considered "clean" as regards identification codes and linkage issues. Documentation, with a list of the cleaning checks and corrections made in the data, was included to enable the NRC to review the cleaning process before the 7th NRC meeting in Vienna in December 2011. A second version of the data was sent once the background data was cleaned and all necessary recoding due to national adaptations was finalized. A third version of the data files was sent to countries when the weights and international achievement scores were available and had been merged with the data files. This version, containing only records that satisfied the sampling standards, 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 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 measures 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, the International Study Center provided item analysis and reliability statistics listing information regarding the number of valid cases, percentages, percentage 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.

Final Product – the TIMSS and PIRLS 2011 International Databases

The data cleaning effort implemented at the IEA DPC ensured that the TIMSS and PIRLS 2011 international databases contained high-quality data (see **2011 TIMSS International Database** and **2011 PIRLS International Database**). 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 2011 International Database and User Guide* and *TIMSS 2011 International Database and User Guide* 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.

References

Foy, P., Drucker, K.T. (Eds.). (2013). *PIRLS 2011 international database and user guide*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Foy, P., Arora, A., & Stanco, G.M. (Eds.). (2013). *TIMSS 2011 international database and user guide*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.