Chapter 8

Creating and Checking the TIMSS 2007 Database

Juliane Barth and Oliver Neuschmidt

8.1 Overview

This chapter describes the TIMSS 2007 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. The overriding concerns were to ensure that all information in the database conformed to the internationally defined data structure, national adaptations to questionnaires were reflected appropriately in the codebooks and documentation, and all variables used for international comparisons were comparable across countries. Quality control measures were applied throughout the process to assure the quality and accuracy of the TIMSS data.

8.2 Steps Taken to Confirm the Integrity of the TIMSS 2007 International Database

The following summarizes the steps taken at all institutions to confirm the integrity of the international database. First, the IEA DPC was responsible for checking the data files from each country, applying standard cleaning rules to verify the accuracy and consistency of the data, and documenting electronically any deviations from the international file structure. Any queries were addressed to the national centers, and modifications were made to the data files as necessary. 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.

When the national files had been checked, the IEA DPC provided national univariate and reliability statistics to the national centers, while the TIMSS & PIRLS International Study Center provided them with data almanacs containing international univariate statistics and national item statistics so that National Research Coordinators (NRCs) could examine their data from an international perspective. This was one of the most important checks in terms of ensuring the international comparability of the data. A particular statistic may seem plausible in a national context, but it may be an outlier when comparing data across countries in an international context. Any such instances were investigated and, if necessary, addressed by either recoding the affected variables or removing them from the international database.

Once verified and in the international file format, the achievement data were sent to the TIMSS & PIRLS International Study Center where basic item statistics were produced and reviewed. At the same time, the IEA DPC sent data files containing information on the participation of schools and students in each country's sample to Statistics Canada. This information, together with data provided by the NRC tracking forms and the software designed to standardize operations and tasks, was used by Statistics Canada to calculate sampling weights, population coverage, and school and student participation rates.[1]

When the review of the item statistics was completed and Statistics Canada finalized the computation of sampling weights, the TIMSS & PIRLS International Study Center conducted the IRT scaling and generated proficiency scores in mathematics and science for each participating student. The scaling methods and procedures are described in Chapter 11. Once the sampling weights and the proficiency scores had been verified at the TIMSS & PIRLS International Study Center, they were sent to the IEA DPC for inclusion in the international database and for distribution to the national centers.

8.3 Data Checking at the IEA Data Processing and Research Center

As described in Chapter 6, each participating country was responsible for entering their TIMSS 2007 data into the appropriate data files and submitting these files to the IEA DPC, where they underwent an exhaustive process of checking and editing—a process known as data cleaning. To facilitate the data cleaning process, countries were requested to provide the IEA DPC with

[1] See Chapter 5 for details about the TIMSS 2007 sampling design.

detailed documentation of their data, in addition to the data files themselves. This data documentation included copies of all original survey tracking forms, copies of the national versions of test booklets and questionnaires, and completing the *Survey Activities Questionnaire,* an Internet-based questionnaire about countries' data collection activities (TIMSS, 2005-2006). To ensure that all national adaptations to the survey instruments were fully documented, countries also were required to submit National Adaptation Forms (NAFs).

Countries also were asked to send the IEA DPC the sample of test booklets selected for double-scoring the constructed-response items (approximately 1,400 booklets per population). The student responses to constructed-response items in these booklets are digitally scanned and preserved for use in the next cycle of TIMSS in 2011, when they will be rescored by TIMSS 2011 scoring staff to monitor consistency in scoring practices between TIMSS 2007 and 2011.

8.3.1 Quality Control in Data Cleaning

TIMSS is a very large and complex study with very demanding standards for data quality. This requires an extensive set of interrelated data checking and cleaning procedures. To ensure that all procedures were conducted in the correct sequence, that no special requirements were overlooked, and that the cleaning process was implemented independently of the persons in charge, the following steps were undertaken:

- Before their use with 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 registered in a specific database. The date of arrival was recorded, along with any specific issues meriting attention.
- The cleaning was organized following strict rules. Deviations from 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 such "manual" correction was logged using a specially developed editing program (SAS-ManCorr), 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.

- Once the 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.
- IEA DPC staff worked closely with the national centers, and at different steps of the cleaning process, countries were provided with the processed data files and accompanying documentation and statistics, allowing them to thoroughly review and correct any inconsistencies detected (see section 8.4).
- 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 in the *TIMSS 2007 User Guide* (Foy & Olson, 2009). Whenever possible, national deviations 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.

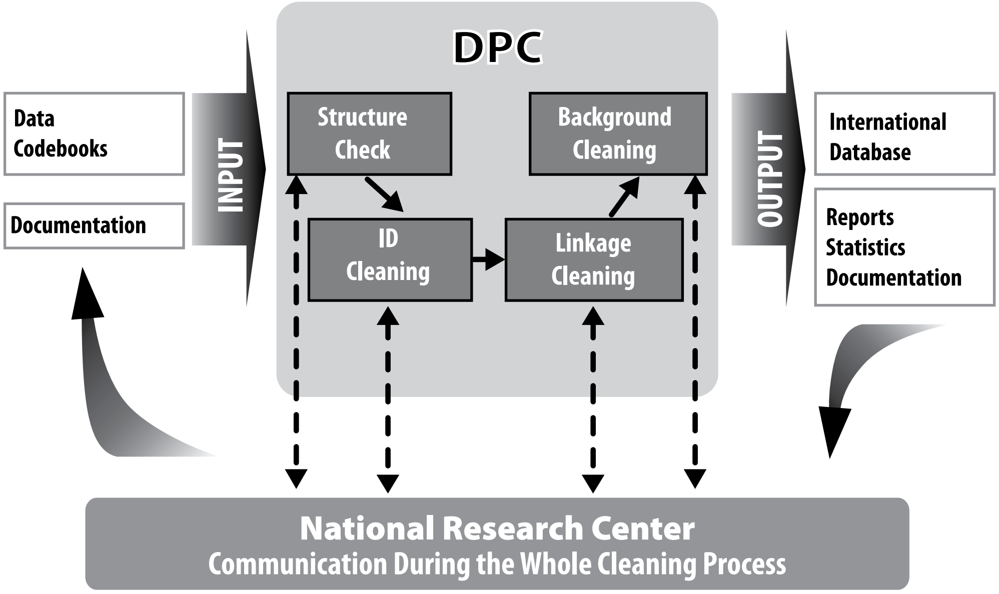
8.3.2 Preparing National Data Files

The main objective of the data cleaning process was to ensure that the data adhered to international formats; school, teacher, and student information could be linked between different survey files; and the data accurately and consistently reflected the information collected within each country.

The program-based data cleaning consisted of the following steps, which are shown in Exhibit 8.1 and explained in the following sections:

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

Exhibit 8.1 Overview of Data Processing at the DPC



8.3.3 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 Adaptation Forms, Student Tracking Forms, Student-Teacher Linkage Forms, Teacher Tracking Forms, and Test Administration Forms. Most countries sent all required documentation along with their data, which greatly facilitated the data checking.

At the beginning of the cleaning process, the tracking information and sampling information captured in the WinW3S database was combined with the WinDEM data files containing the corresponding survey instrument data (see Chapter 6 for more information).

The first checks implemented at the IEA DPC looked for differences between the international file structure and the national file structures. Some countries made adaptations (such as adding national variables or omitting or modifying international variables) to their background questionnaires. 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 National Adaptation Forms as they adapted the international codebooks. Where necessary, the IEA DPC modified the structure of the country's data to ensure that the resulting data remained comparable between countries.

As part of this standardization process, since direct correspondence between the data collection 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 software 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, and included a description of the problem and the action taken by the program or by the IEA DPC staff.

Where problems could not be rectified automatically, they were reported to the responsible NRC so that the original data collection instruments and tracking forms could be checked to trace the source of the errors. Wherever possible, staff at the IEA DPC suggested a remedy and asked the NRCs to either accept it or propose an alternative. Data files then were updated to reflect the solutions agreed upon. Where the NRC could not solve problems by inspecting the instruments or forms, a general cleaning rule was applied to the files to rectify this. After all automatic updates had been applied, remaining corrections to the data files were applied directly by keyboard, using a specially developed editing program (SAS-ManCorr).

8.3.4 Identification Variable (ID) Cleaning

Each record in a data file should have a unique identification number. The existence of records with duplicate ID numbers in a file implies an error of some kind. If two records share the same ID number, 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 numbers and it was impossible to identify which record contained the "true data," both records were removed from the database. The IEA DPC tried to keep such losses at a minimum, and in only a few cases were data actually deleted.

The ID cleaning focused on the student background questionnaire file, which contained most of the critical ID variables. Apart from the unique student ID number, variables pertaining to the student participation and exclusion status, as well as the dates of birth and dates of testing used to calculate age at the time of testing were important to check. The Student Tracking Forms[2] were essential in resolving any anomalies, as was close cooperation with NRCs (since, in most cases, the Student Tracking Forms were completed in the country's official language). After cleaning, databases created from the WinW3S program containing information about student participation and exclusion were sent to Statistics Canada, where they were used to calculate students' participation rates, exclusion rates, and student sampling weights.

8.3.5 Linkage Check

In TIMSS, data about students and their schools and teachers appeared in several different files, so that it was crucial that the records from these files link together correctly to provide meaningful data for analysis and reporting. The linkage was implemented through a hierarchical ID numbering system incorporating a school, class, and student component[3] and was cross-checked against the tracking forms. It was necessary that students' entries in the achievement file and student background file were matched correctly; that the student entries in the reliability scoring file matched of the student entries in the achievement file; that the teachers were linked to the correct students; and that the schools were linked to the correct teachers and students.

8.3.6 Resolving Inconsistencies in Background Questionnaire Data

The number of inconsistent and implausible responses in background files varied from country-to-country, but no country's data were 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. All background questionnaire data were checked for consistency among the responses given. For example, question number 1(a) in the *School Questionnaire* asked for the total school enrollment (number of students) in all grades, while 1(b) asked for the enrollment in the target grade only. Clearly, the number given for 1(b) should not exceed the number given for 1(a). All such inconsistencies that were detected were flagged and

[2] Tracking forms were used to record the sampling of schools, classes, teachers, and students (also see Chapter 6).

[3] The ID number of a higher level is included in the ID number of a lower sampling level. The class ID includes the school ID, and the student ID includes the class ID (e.g., student 1220523 may be described as student 23 of class 05 in school 122).

the NRCs asked to investigate. Those cases that could not be corrected or where the data made no sense were recoded to "Omitted".

Filter questions, which appear in some questionnaires, were used to direct the respondent to a particular section of the questionnaire. Filter questions and the dependent questions that follow were subject to the following cleaning rules: If the answer to the filter question was "No" or "Not applicable" and the dependent questions were answered, then the filter question was recoded to "Yes".

Split variable checks were applied to questions where the answer was coded into several variables. For example, question 5 in the *Student Questionnaire* listed a number of home possessions and asked the student to check all that applied. Student responses were captured in a series of nine variables, each one coded as "Yes" if the corresponding possession was checked and "No" if left unchecked. Occasionally, students checked the "Yes" boxes but left the "No" boxes unchecked or missing. Since in these cases, it was clear that the unchecked boxes actually meant "No," these were recoded accordingly.

8.3.7 National Cleaning Documentation

NRCs received a detailed report (IEA, 2007) of all problems identified in their data and the steps applied to correct them. These included the following:

- Documentation of any data problems detected by the cleaning program and the steps applied to resolve them
- 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 scores in mathematics and science that could be used in national analyses to be conducted before the international database became available.

8.3.8 Handling of Missing Data

When the TIMSS 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. 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. This code also was used for responses that were not interpretable in both the background and the achievement files (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).

8.4 Data Products

Data products sent to NRCs by the IEA DPC and the TIMSS & PIRLS International Study Center included both data almanacs and data files.

8.4.1 Data Almanacs and Item Statistics

Each country received a set of data almanacs or summaries, 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. The data almanacs were sent to participating countries for review. When necessary, they were accompanied by specific questions about the data presented in them. They also were used by the TIMSS & PIRLS International Study Center during the data review and in the production of the reporting exhibits. Also, the IEA DPC produced a set of preliminary scoring reliability statistics for each country containing summary statistics at the item level on the percent of agreement between scorers.

8.4.2 Versions of the National Data Files

Building the international database was an iterative process. The IEA DPC provided each NRC with a new version of their country's data files whenever a major step in data processing was completed. This also guaranteed that NRCs had a chance to review their data and run their own checks to validate the data files. Before the TIMSS international database was published, several 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 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 NRC to review the cleaning process. Another version of the data files was sent to countries when the weights and international achievement scores were available and had been merged in the files, together with the data almanacs. This was done after all exhibits of the TIMSS international reports had been verified and final updates to the data files implemented, and enabled the NRCs to replicate the results presented in the international reports.

8.4.3 The TIMSS 2007 International Database

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

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

In a joint effort of the IEA DPC and the TIMSS & PIRLS International Study Center, a National Adaptations Database containing all adaptations to questionnaires made by individual countries and documenting how they were handled was constructed. The meaning of country-specific items also can be found in this database, as well as recoding requirements by the TIMSS & PIRLS International Study Center. Information contained in this database is provided in the *TIMSS 2007 User Guide for the International Database* (Foy & Olson, 2009) upon release of the TIMSS 2007 data.

References

Foy, P. & Olson, J.F. (2009). *TIMSS 2007 user guide for the international database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

IEA. (2007). *General cleaning documentation V11*. Hamburg: IEA Data Processing and Research Center.

TIMSS (2005-2006). *TIMSS 2007 survey operations procedures, units 1-6.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.