Chapter 7

Creating and Checking the TIMSS Advanced 2008 Database

Milena Taneva

7.1 Introduction

Creating the TIMSS Advanced 2008 database and ensuring its integrity was a complex endeavor requiring close coordination and cooperation among the staff at the IEA Data Processing and Research Center (IEA DPC), the TIMSS & PIRLS International Study Center at Boston College, 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 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 TIMSS data. This chapter describes the data checking and database creation procedures implemented by the IEA DPC in collaboration with the TIMSS &

PIRLS International Study Center and Statistics Canada, and the steps taken to confirm the integrity of the international database.

7.2 Overview of International Database Construction

On receipt of the data files from each country, the IEA DPC assumed responsibility for checking them for completeness, for applying standard cleaning rules to verify the accuracy and consistency of the data, and for documenting any deviation from the international file structure. All queries were communicated to the national centers and modifications were made to the data files as necessary. After all modifications had been applied, the data were processed and checked again. This process of editing the data, checking the edit reports, and implementing corrections was repeated as many times as necessary until all data were consistent and comparable nationally and internationally.

In preparation for creating the international database, the TIMSS & PIRLS International Study Center provided countries with data almanacs containing international statistics summarizing the responses to the questions in the background questionnaires as well as national statistics summarizing all achievement items so that National Research Coordinators (NRCs) could examine their data in an international perspective. This was one of the most important checks toward international comparability of the data. While in a national context a particular response may seem plausible, it may raise issues when viewed in an international context. All such issues were addressed, and the corresponding variables were either recoded or removed from the international database, as appropriate. Once the achievement data had been verified and converted to the international file format, they 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 NRCs from tracking forms and the IEA's sampling software, was used by Statistics Canada to calculate sampling weights, population coverage and exclusion statistics, and participation rates.[1]

When the item review was completed and the computation of sampling weights was finalized and verified, the TIMSS & PIRLS International Study Center proceeded to scale the TIMSS Advanced 2008 achievement data and produce proficiency scores in advanced mathematics and physics for all participating students.[2] Once 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.

7.3 Software Support for National Centers

The IEA DPC went to great lengths to ensure that the data received from all participating countries were of high quality and internationally comparable. The foundations for quality assurance were laid before the first data arrived at the IEA DPC through the provision to the participants of software designed to standardize a range of operational and data-related tasks.

The WinW3S software (IEA, 2007) performed all within-school sampling operations adhering strictly to the sampling rules established by Statistics Canada and the TIMSS & PIRLS International Study Center. The software also created all necessary tracking forms and stored student and teacher tracking information—such as students' age, gender, and participation status.

The WinDEM program (IEA, 2008) enabled key-entry of all TIMSS Advanced 2008 achievement and questionnaire data in a

1 Sampling weights and participation rates are described in Chapter 4.

2 The item review process and the scaling of the TIMSS Advanced 2008 achievement data are described in Chapter 8.

standard, internationally defined format. The software also included a range of checks for data verification within and across the various data files. These checks were applied by national center staff before sending the files to the IEA DPC.

7.4 Checking the Data at the IEA DPC

All participating countries were responsible for entering their TIMSS Advanced 2008 data into the appropriate data files and submitting them 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 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 version of the assessment booklets and questionnaires, and the completed Survey Activities Questionnaire. To ensure that all national adaptations to the survey instruments were fully documented, countries also were required to submit National Adaptations Forms, which became a written record of all national adaptations.

Countries also were asked to send the IEA DPC all test booklets selected for double-scoring the constructed-response items (some 500 advanced mathematics booklets and 500 physics booklets per country). The students' responses to constructed-response items in these booklets were digitally scanned and preserved for use in the next cycle of TIMSS Advanced, when they will be rescored by the national scoring staff to monitor consistency in scoring practices across survey cycles.

7.4.1 Quality Control in Data Cleaning

TIMSS Advanced is a complex study with demanding standards for data quality. This required 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 being applied to real data, all data-cleaning programs were thoroughly tested using simulated data sets containing all imaginable problems and inconsistencies.
- All incoming data and documents from countries were registered in a database. The date of arrival was recorded, along with any specific issues that merited attention.
- The data cleaning was organized following strict rules. Deviations from the cleaning sequence were not allowed, and the scope for involuntary changes to the cleaning procedures was minimal.
- All corrections applied to a country's data files were documented in a country-specific cleaning report.
- Occasionally, it was necessary to make manual changes to a country's data files. All manual corrections were documented in a program 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 data cleaning.
- Once data-cleaning was completed for a country, all cleaning steps were repeated from the beginning with the cleaned data files to detect any problems that might have been inadvertently introduced during the cleaning process.
- IEA DPC staff worked closely with the national research centers and, at different steps of the cleaning process, countries were provided with the processed data files with accompanying documentation and statistics to review and correct any inconsistencies detected.

- 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 were recorded in a national adaptations database for the *TIMSS Advanced 2008 User Guide for the International Database* (Foy & Arora, 2009). All national deviations required recoding of the data to comply with the international data structure. However, if international comparability could not be assured, the problematic data were removed from the international database.

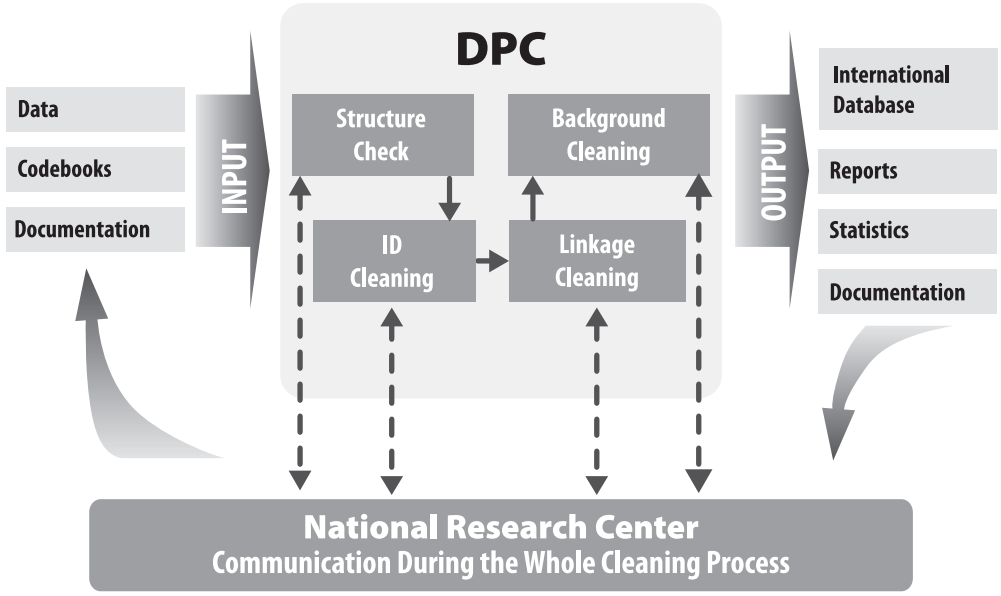
7.4.2 Preparing National Data Files

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

- Structure check and documentation
- Identification variable cleaning
- Linkage check
- Resolving inconsistencies in background questionnaire data

These data cleaning steps are illustrated in Exhibit 7.1 and are described in the subsequent sections. As shown in the exhibit, national research centers submitted their data and documentation to the IEA DPC. These data were cleaned by the DPC, which then produced the international database and provided documentation of the cleaning process to the national centers, including reports and summary statistics. Throughout the data cleaning process, effective communication between the IEA DPC and the national centers was key to ensuring quality data in the end.

Exhibit 7.1 Overview of Data Processing at the DPC



7.4.3 Structure Check and Documentation

For every country, data cleaning began with an exploration of its data file structure and a review of its documentation: National Adaptations Forms, Student Tracking Forms, Student-Teacher Linkage Forms, Teacher Tracking Forms, and Test Administration Forms (see Chapter 5 for an overview of the TIMSS Advanced data collection forms). Most countries sent all required documentation along with their data, which greatly facilitated the data checking.

At the beginning of the cleaning process, tracking and sampling information captured in the WinW3S database was combined with the actual data files that contained the survey data.

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 to their background

questionnaires (such as adding national variables, or omitting or modifying international variables). The extent and nature of these changes differed across countries. Some countries administered the questionnaires without any changes (apart from translation), while other countries inserted items or options within existing international variables or added entirely new national questions. To keep track of all adaptations, NRCs were asked to complete National Adaptations Forms that became a written record of all implemented national adaptations. Where necessary, the IEA DPC modified the structure of a country's data to ensure that the resulting data remained comparable with those of other countries.

Once all data files matched the international standard as specified in the international codebooks,[3] a series of standard cleaning rules were applied to the data. This was conducted using software developed at the IEA DPC that could identify and, in many cases, automatically correct inconsistencies in the data. Each problem was recorded in a database, identified by a unique problem number and with a description of the problem and the action taken by the program or by the IEA DPC staff.

When problems could not be resolved automatically, they were reported to the national centers so that original data collection instruments and tracking forms could be checked to trace the source of the errors. Whenever possible, staff at the IEA DPC suggested solutions and asked the national centers to either accept them or propose alternatives. Data files then were updated to reflect the solutions agreed upon. When the national centers could not solve problems, they were corrected by applying generalized cleaning rules.

As part of this standardization process, since a 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-

3 The codebooks defined the structure of the data files and are described in Chapter 5.

oriented layout more suited to data analysis. Variables created purely for verification purposes during data entry were dropped at this time and provision was made for additional variables necessary for analysis and reporting (i.e., reporting variables, derived variables, sampling weights, and achievement scores).

7.4.4 Identification Variable Cleaning

Each record in a data file should have a unique identification (ID) number. The existence of records with duplicate ID numbers in a file implies an error of some kind. When two records shared the same ID number and contained exactly the same data, one of the records was deleted and the other remained in the database. When records contained different data, apart from the ID numbers, and it was not possible to identify which record contained the "true data," both records were removed from the database. The IEA DPC tried to keep such losses to a minimum, and only in rare cases were data actually deleted.

The ID cleaning focused on the student background questionnaire files, which contained most of the critical identification variables. Apart from the unique student ID numbers, variables pertaining to the students' participation and exclusion status as well as the dates of birth and dates of testing used to calculate students' age at the time of testing were important to check. The Student Tracking Forms were essential in resolving all anomalies, as was close cooperation with the national centers (since, in most cases, the Student Tracking Forms were completed in each country's national language). Once ID cleaning was completed, the WinW3S databases with information about student participation or exclusion were sent to Statistics Canada where they were used to calculate sampling weights, exclusion rates, and participation rates.

7.4.5 Linkage Check

In TIMSS Advanced 2008, data about students and their schools and teachers were located in several different files, so that it was crucial that the records from these files linked together correctly to provide meaningful data for analysis and reporting. The linkage across files was implemented through a hierarchical ID numbering system that incorporated a school, class, and student component[4] and was cross-checked against the tracking forms. It was required that student records in the achievement files and student background files be matched correctly, that student entries in the reliability files be properly matched to student entries in the achievement files, that teachers be linked to the correct students, and that schools be linked to the correct teachers and students.

7.4.6 Resolving Inconsistencies in Background Questionnaire Data

The number of inconsistent or implausible responses in background files varied from country to country, but no country's data were completely free of inconsistent responses. Treatment of inconsistent responses was determined on a question-by-question basis, using available documentation to make informed decisions. All background questionnaire data were checked for consistency among the responses given. For example, question 1a in the school questionnaire asked for total school enrollment in all grades, while 1b asked for enrollment in the target grade only. Clearly, the response given to 1b should not exceed the response given to 1a. All such inconsistencies were flagged, and the national centers were asked to investigate. In cases that could not be resolved, or where the data made no sense, responses were recoded as "Omitted".

Filter questions, which appear in some questionnaires, were used to direct respondents to a particular section of the questionnaire. Filter

4 The Hierarchical ID numbering system consisted of a school ID, a class ID that included the school ID, and a student ID that included the class ID, such that student 01220523 could be identified as student 23 in class 05 of school 0122.

questions and the dependent questions that followed were subject to specific cleaning rules. If the answer to a filter question was "No", or simply not given, and yet the dependent questions were answered, then the filter question was recoded as "Yes".

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

7.4.7 National Cleaning Documentation

NRCs received detailed reports of all problems identified in their data and of any steps applied to correct them, which also included records of all deviations from the international version of the 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 national standardized scores of advanced mathematics and physics that could be used in national analyses to be conducted before the official release of the international database.

7.4.8 Handling Missing Data

When the TIMSS Advanced 2008 data were entered with WinDEM, two types of entries were possible: valid data values and missing

data values. Missing data could 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 and consequently he or she had no opportunity to provide an answer to the question.
- Omitted: The respondent had a chance to answer the question, but chose not to. This code also was used for responses that were not interpretable in either the background files or the achievement files.
- Logically not applicable (applied to filter-dependent questions): The respondent answered a filter question in a way that made the following dependent questions not applicable to him or her.
- Not reached (used only in the achievement files): An item was considered not reached when the item itself and the one immediately preceding it were not answered, and there were no other items completed in the remainder of the assessment booklet.

7.5 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.

7.5.1 Data Almanacs and Item Statistics

Every country received a set of data almanacs produced by the TIMSS & PIRLS International Study Center. They contained weighted national summary statistics for each variable included in the survey instruments and were sent to the participating countries for review. When necessary, they were accompanied by specific questions about the data presented

in them. Countries also were provided item almanacs with national summary statistics for the achievement items, which were reviewed by the national centers as part of the data validation process.

7.5.2 Versions of the National Data Files

Building the TIMSS Advanced 2008 international database was an iterative process. The IEA DPC provided NRCs with an updated version of their country's data files whenever a major step in data processing was completed. This guaranteed that the NRCs had a chance to review their data and run their own checks to validate the contents of the data files. Prior to the release of the TIMSS Advanced 2008 international database, the participating countries received several versions of their national data files. Each country received only its own data. The first version was sent as soon as the data could be regarded as "clean" with respect to identification variables and any possible linkage issues. These first files contained interim national 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 to the data, also was sent to enable NRCs to review the cleaning process. Another version of the data files was sent to the countries when the sampling weights and the international achievement scores were available. This was done after all exhibits of the TIMSS Advanced 2008 international report had been verified and final updates to the data files implemented. This allowed NRCs to replicate the results presented in the international report.

7.5.3 The TIMSS Advanced 2008 International Database

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

- Data recorded for each variable were 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 respondents—students, teachers, and school principals.
- Sampling weights and student achievement scores were available for international comparisons.

In a joint effort involving the IEA DPC and the TIMSS & PIRLS International Study Center, a national adaptations database, describing all adaptations made by individual countries to questionnaires and how they were handled, was constructed. All information contained in this database is provided in Supplement 2 of the *TIMSS Advanced 2008 User Guide*.

References

Foy, P., & Arora, A. (2009). *TIMSS Advanced 2008 user guide for the international database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

IEA (2007). *Windows within-school sampling software (WinW3S)* [Computer software and manual]. Hamburg: IEA Data Processing and Research Center.

IEA (2008). *Windows data entry manager software (WinDEM)* [Computer software and manual]. Hamburg: IEA Data Processing and Research Center.


OCCASVS
REPVSCVLVM