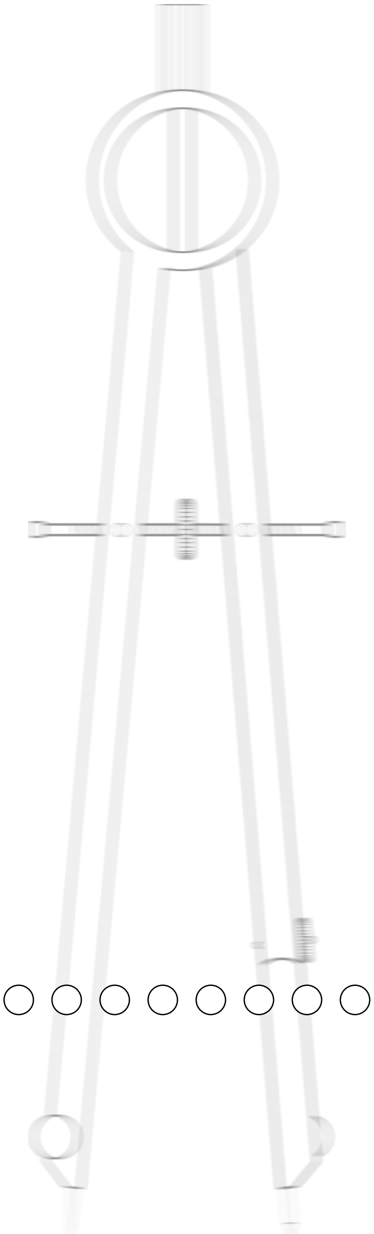
10

Data Management and Database Construction

Dirk Hastedt
Eugenio J. Gonzalez



10 Data Management and Database Construction

Dirk Hastedt
Eugenio J. Gonzalez

10.1 Overview

The TIMSS 1999 data were processed in close cooperation among the TIMSS International Study Center at Boston College, the IEA Data Processing Center, the Educational Testing Service, Statistics Canada, and the national research centers of the participating countries. Under the direction of the International Study Center, each institution was responsible for specific aspects of the data processing.

Data processing consisted of six general tasks: data entry, creation of the international database, calculation of sampling weights, scaling of achievement data, analysis of the background data, and creation of the exhibits for the international reports. Each task was crucial to ensuring the quality and accuracy of the TIMSS results. This chapter describes the data entry task undertaken by each national research coordinator,[1] the data checking and database creation that was implemented by the IEA Data Processing Center, and the steps taken to ensure the quality and accuracy of the international database.[2] It discusses the responsibilities of each participant in creating the international database; the flow of the data files among the centers involved in the data processing; the structure of the data files submitted by each country for processing, and the resulting files that are part of the international database; the rules, methods, and procedures used for data verification and manipulation; the data products created during data cleaning and provided to the national centers; and the computer software used in that process.

10.2 Data Flow

The data collected in the TIMSS 1999 survey were entered into data files with a common international format at the national research centers of the participating countries. These data files were then submitted to the IEA Data Processing Center for clean-

[1] Further information about the role of the national research coordinator in entering and checking the TIMSS data may be found in Chapter 7, TIMSS Field Operations.

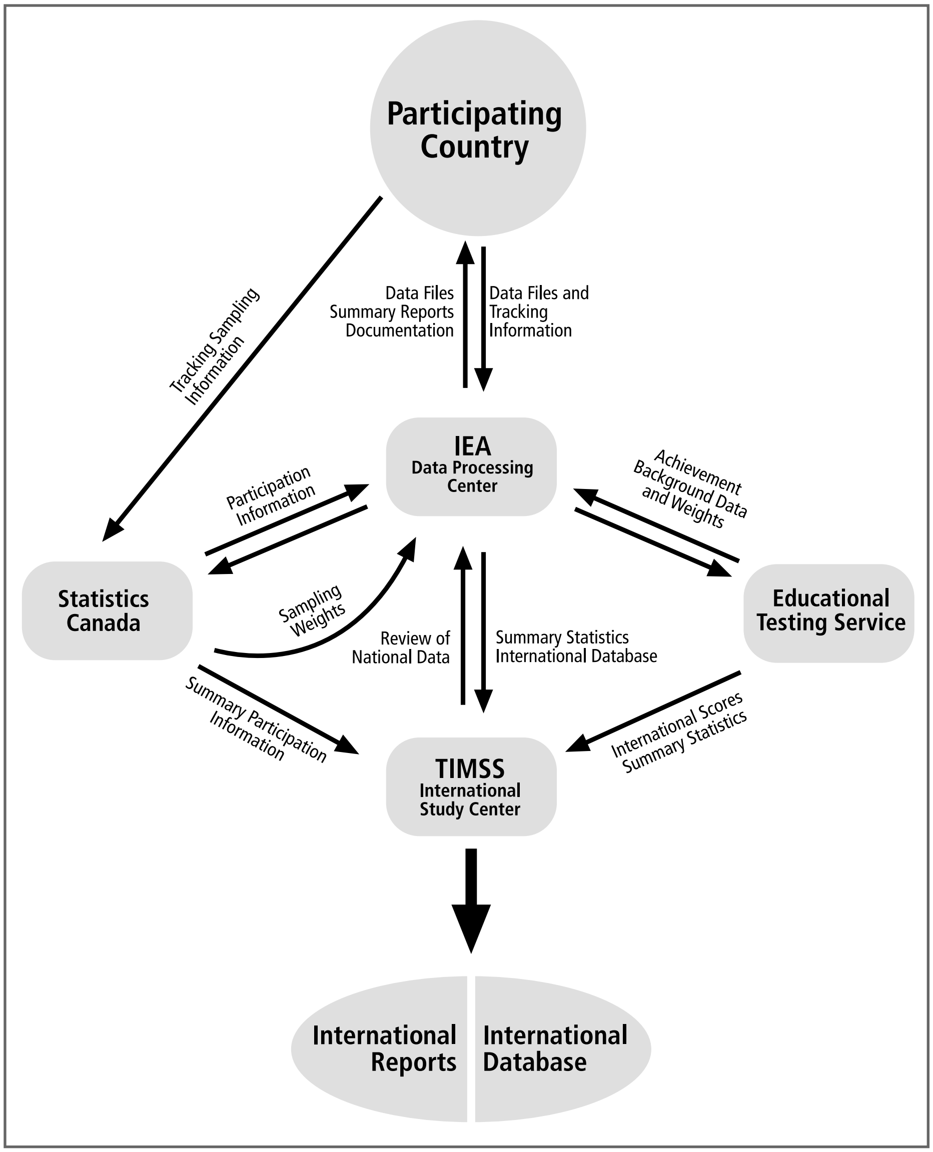
[2] The TIMSS weighting, scaling, and analysis procedures are described in Chapters 11, 14, and 15, respectively.

ing and verification. The major responsibilities of the Data Processing Center were to check that the data files received matched the international standard and to make modifications where necessary; to apply standard cleaning rules to the data to verify their consistency and accuracy; to interact with the national research coordinators (NRCs) to ensure their accuracy; produce summary statistics of the background and achievement data for review by the TIMSS International Study Center; and finally, upon feedback from the individual countries and the TIMSS International Study Center, to construct the international database. The IEA Data Processing Center was also responsible for distributing the national data files to each of the participating countries.

Once verified and in the international file format, the achievement data were sent to the International Study Center where basic item statistics were produced and reviewed.[3] At the same time the Data Processing Center 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 national research coordinator from tracking forms and the within-school sampling software, was used by Statistics Canada to calculate sampling weights, population coverage, and school and student participation rates. The sampling weights were sent to the TIMSS International Study Center for verification and then forwarded to Educational Testing Service for use in scaling the achievement data. When the review of the item statistics was completed and the Data Processing Center had updated the database, the student achievement files were sent to Educational Testing Service who conducted the scaling and generated proficiency scores in mathematics and science for each participating student. Once the sampling weights and the proficiency scores were verified at the International Study Center, they were sent to the Data Processing Center for inclusion in the international database and then distributed to the national research centers. The International Study Center prepared the exhibits for the international reports and published the results of the study. Exhibit 10.1 is a pictorial representation of the flow of the data files.

[3] The item review process is described in Chapter 13.

Exhibit 10.1 Flow of Data Files


Participating Country
Tracking Sampling Information
Data Files
Summary Reports
Documentation
Data Files and Tracking Information
IEA Data Processing Center
Participation Information
Achievement Background Data and Weights
Statistics Canada
Sampling Weights
Review of National Data
Summary Statistics International Database
Educational Testing Service
Summary Participation Information
International Scores Summary Statistics
TIMSS International Study Center
International Reports
International Database

10.3 Data Entry at the National Research Centers

Each TIMSS 1999 national research center was responsible for transcribing the information from the achievement booklets and questionnaires into computer data files. Data-entry software, DataEntryManager (DEM), adapted specifically for TIMSS 1999 was provided to each participating country, together with the *Manual for Entering the TIMSS-R Data* (TIMSS, 1998) and codebooks describing the layout of the data. The codebooks contained information about the variable names used for each variable in the survey instruments, and about field length, field location, labels, valid ranges, default values, and missing codes.

The codebooks were designed to be an integral part of the DEM system for data entry, although they could also be used in conjunction with other data-entry systems. Although DEM was the recommended software, some of the participating countries elected to use a different data entry system. Data files were accepted from the countries provided they conformed to the specifications given in the international codebooks.

In order to facilitate data entry, the codebooks and data files were structured to match the tests and questionnaires. This meant that there was a data file for each survey instrument. Each country was responsible for submitting six data files: Student Background, Student Achievement, Scoring Reliability, Mathematics Teacher Background, Science Teacher Background, and School Background. Each file had its own codebook.

The following files were used during data entry

- The Student Background Data File contained one record for every student listed on the *Student Tracking Form* for each sampled class, including students that did not participate in the testing session and students that were excluded. There are two versions of the student questionnaire, one tailored to countries where the sciences are taught as a general integrated subject and the other for countries where biology, physics, and chemistry are taught as separate subjects. Each version of the questionnaire had its own codebook and data file layout. The data for the student background data file came from the *Student Tracking Form*, the *Teacher Student Linkage Form*, and the *Test Administration Form*, as well as from the *Student Background Questionnaire*.
- The Student Achievement Data File contained a record for each student that completed an achievement test booklet. Each record contains the student's responses to whichever of the 8 test booklets was assigned to the student.
- In order to check the reliability of the free-response item scoring, the free-response items in a random sample of 25 percent of booklets were scored independently by a second scorer. The responses from these booklets were stored in a Scoring Reliability File, which contained one record for each student in the reliability sample.

- The Teacher Questionnaire Data Files contained one record for every entry on the *Teacher Tracking Form*, including teachers who did not complete a teacher questionnaire. There were separate data files for mathematics and science teachers. The data for this file came from the mathematics and science teacher questionnaires and from the teacher tracking form.
- The School Data file contained one record for each school sampled to take part in the study, whether the school participated or not. The file also contained a record for each replacement school in which data collection took place. The data for this came from the school questionnaire and the school tracking form.

10.3.1 Data Files Received by the IEA Data Processing Center

In addition to the data files, countries were also required to submit documentation supporting their field procedures and copies of their national translated tests and questionnaires. The documentation included a report of their survey activities, a series of data management forms with clear indications of any changes made in the tests or the layout of the data files, and copies of all sampling and tracking forms. These materials were archived at the IEA Data Processing Center and kept for reference during data processing.

Each country was provided with a program called LINKCHK to carry out checks on the data files before submitting them to the IEA Data Processing Center. The program was designed to help NRCs perform an initial check of the system of student, teacher, and school identification numbers after data entry, both within and across files. The reports produced by the LINKCHK program allowed countries to correct problems in the identification system before transferring the data to the IEA Data Processing Center.

10.4 Data Cleaning at the IEA Data Processing Center

Once the data were entered into data files at the national research center, the data files were submitted to the IEA Data Processing Center for checking and input into the international database. This process is generally referred to as data cleaning. The goals of the TIMSS 1999 data cleaning were to identify, document, and, where necessary and possible, correct deviations from the international file structure, and to correct key punch errors, systematic deviations from the international data formats, problems in linking observations across files, inconsistent tracking information across and within files, and inconsistencies within and across observations. The main objective of the process was to ensure that the data adhered to international formats and accurately and consistently reflected the information collected within each country.

Data cleaning involved several steps. Some of these were repeated until satisfactory results were achieved. During the first step, all incoming data files were checked and reformatted as necessary so that their file structure conformed to the international format. As a second step, all problems with identification variables, linkage across files, codes used for different groups of variables, and participation status were detected and corrected. The distribution for each variable was examined, with particular attention to variables that presented implausible or inconsistent distributions based on the information from the country involved and on other answers in the questionnaires.

During this stage, a series of data summary reports were generated for each country. The reports contained listings of codes used for each variable and pointed to outliers and changes in the structure of the data file. They also contained univariate statistics. The reports were sent to each participating country, and the NRC was asked to review the data and advise how best to resolve inconsistencies. In many cases the NRC was asked to go back to the original booklets from which the data had been entered.

In data cleaning two main procedures were used to make necessary changes in the data. Inconsistencies that could unambiguously be solved were corrected automatically by a program using standard cleaning routines. Errors that could not be solved with these routines had to be solved case by case by the DPC staff. In either case, all changes made in the data were documented. A database was created in which each change was recorded, and it was possible to reconstruct the original database received from a country.

10.4.1 Standardization of the National File Structure

The first step in the data processing at the international level was to verify the compatibility of the national datasets with the international file structure as defined in the TIMSS 1999 international codebook. This was necessary before the standard cleaning with the Data Processing Center cleaning software could be performed.

Although the TIMSS 1999 international codebooks distributed with the data entry software gave clear and detailed instructions about the structure and format of the files each country was to submit to the IEA Data Processing Center, some countries opted to enter and submit their data files in other formats, using structures different from the international standard. For the most part, these differences were due to specific national circumstances. The manual for entering the TIMSS 1999 data asked countries to prepare and send their data files using the DEM software, which produces an extended dBase format file. Some data files, however, were received in ASCII fixed format (raw data), SPSS format, and SAS format.

After the national files were converted into dBase format, the structure of the files was inspected and deviations from the international file structure were identified. A custom-designed program scanned the file structure of the files for each country and identified the following deviations:

- International variables omitted
- National variables added
- Different variable length or number of decimal positions
- Different coding schemes or out-of-range values
- Specific national variables
- Gang-punched variables

Together with the inspection of the national data files, the data management and tracking forms submitted by each NRC were reviewed. As a result of this initial review, the Data Processing Center outlined and implemented necessary changes in the national data to make the files compatible with the international format. In most cases programs had to be prepared to fit the file structures and particularities of each country.

As part of the standardization process, the file structure was rearranged to facilitate data analysis, since direct correspondence between the files and the data-collection instruments was no longer necessary. At this stage also, the Student Background file and Student Achievement files were merged to form a single student file. Variables created during data entry for verification only were omitted from all files at this time, and new variables were added (i.e., reporting variables, derived variables, sampling weights, and achievement scores).

10.4.2 Cleaning Rules and Procedures

After the data files received from the countries were transformed into the international format, certain standard checks and cleaning rules were applied to each file.

The first step was to check for deviations from international standards in both data-collection instruments and data files. Instruments were checked for missing questions or items, changes in the number of answer categories, alterations in coding schemes, and other national adaptations. Data files were examined for missing variables, changes in field length and number of decimal places, modifications of coding schemes, and additional national variables.

After all deviations from the international standard had been identified, a cleaning program was run on each file to make a set of standard changes. This was to facilitate the application of more specific cleaning rules at the next stage. After this step, each data file matched the international standard as specified in the international codebook. Among changes made at this time were adjustments to the hierarchical identification number system, differentiation between 'Not Applicable', 'Missing', and 'Not Administered' codes, adding omitted variables and coding them as 'Not administered', and recoding systematic deviations from the international coding scheme.

The TIMSS 1999 cleaning program labelled each problem with an identification number, a description of the problem, and the action taken by the program. All problems were recorded in an error database containing one error file for each file that was checked. As problems were identified that could not be automatically rectified, they were reported to the responsible NRC so that data-collection instruments and tracking forms could be checked to trace the source of the errors. Wherever possible, staff at the IEA Data Processing Center suggested a remedy, and asked the NRCs to either accept it or propose an alternative. The data files were updated as solutions to problems were found. Where problems could not be solved by the NRC by inspecting the instruments or forms, they were rectified by applying a general cleaning program.

After all automatic updates had been applied, remaining corrections to the data files were applied directly, or manually, using a specially developed editing program. The manual corrections took into account country-specific information that could not be used by the cleaning program.

10.4.3 National Cleaning Documentation

National research coordinators received a detailed report of all problems identified in their data, and of the steps taken to correct them. These included:

- A record of all deviations from the international data-collection instruments and the international file structure
- Documentation of the data problems uncovered by the cleaning program and the steps taken to resolve them
- A list of all manual corrections made in each data file

In addition to documentation of data errors and updates, the IEA Data Processing Center provided each NRC with new data files that incorporated all agreed updates. The data files were transformed from the standard layout designed to facilitate data entry to a new format oriented more toward data analysis. The updated 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.

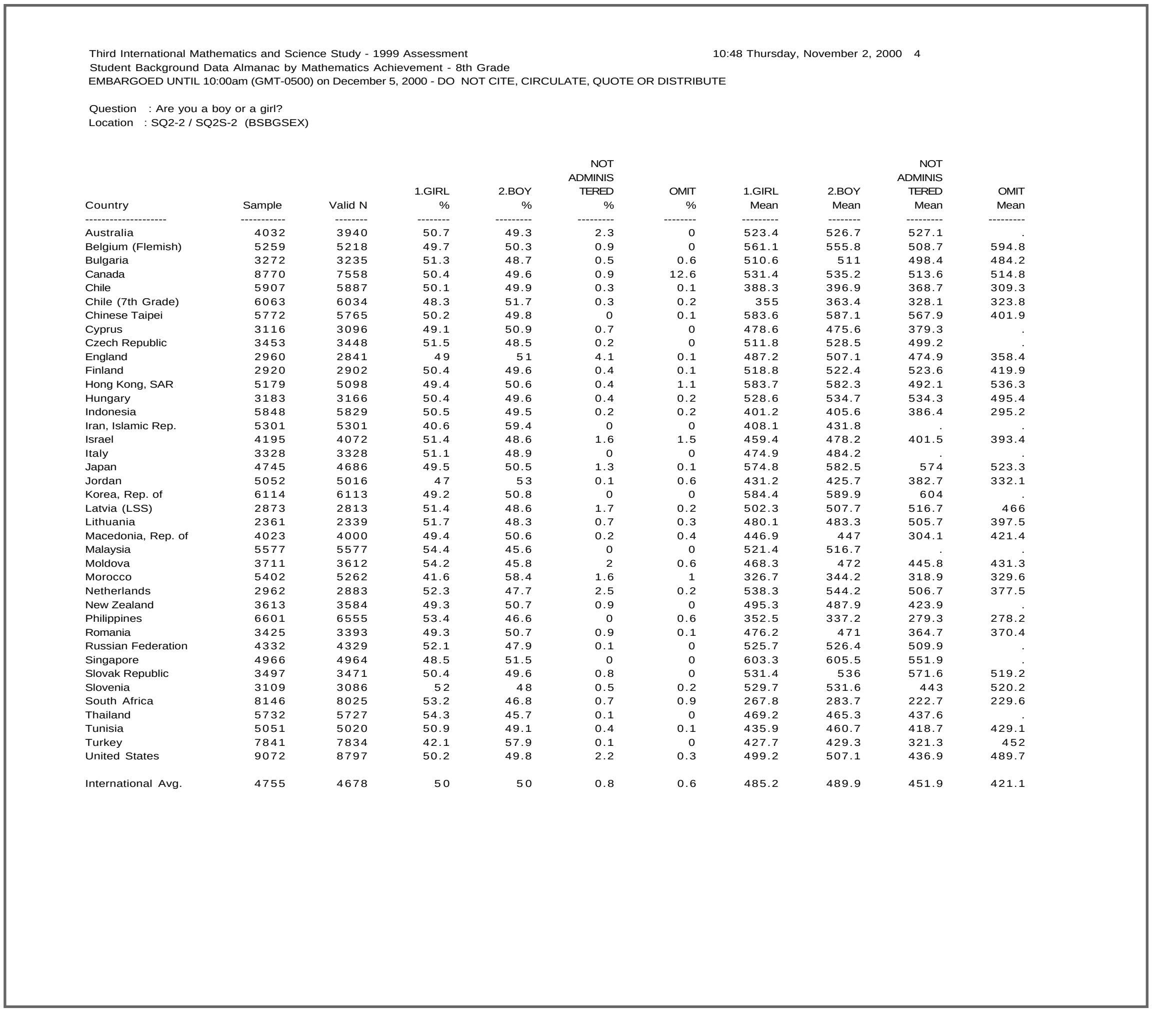
10.5 Data Products

Data products sent by the IEA Data Processing Center to NRCs included both data almanacs and data files.

10.5.1 Data Almanacs

Each country received a set of data almanacs, or summaries, produced by the TIMSS International Study Center. These contained weighted summary statistics for each participating country on each variable included in the survey instruments. There were two types of display. The display for categorical variables included an estimate of the size of the student population, the sample size, the weighted percentage of students who were not administered the question, the percentage of students choosing each of the options on the question, and the percentage of students who chose none of the valid options. The percentage of students to whom the question did not apply was also presented. For continuous variables the display included an estimate of the size of the student population, the sample size, the weighted percentage of students who were not administered the question, the percentage who did not respond, the percentage to whom the question did not apply, the mean, mode, minimum, maximum, and the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. An example of such data displays is presented in Exhibits 10.2 and 10.3. These data almanacs were sent to the participating countries for review. When necessary, they were accompanied by specific questions about the data presented in them. They were also used by the TIMSS International Study Center during the data review and in the production of the reporting exhibits.

Exhibit 10.2 **Example Data Almanac Display for a Categorical Student Background Variable**



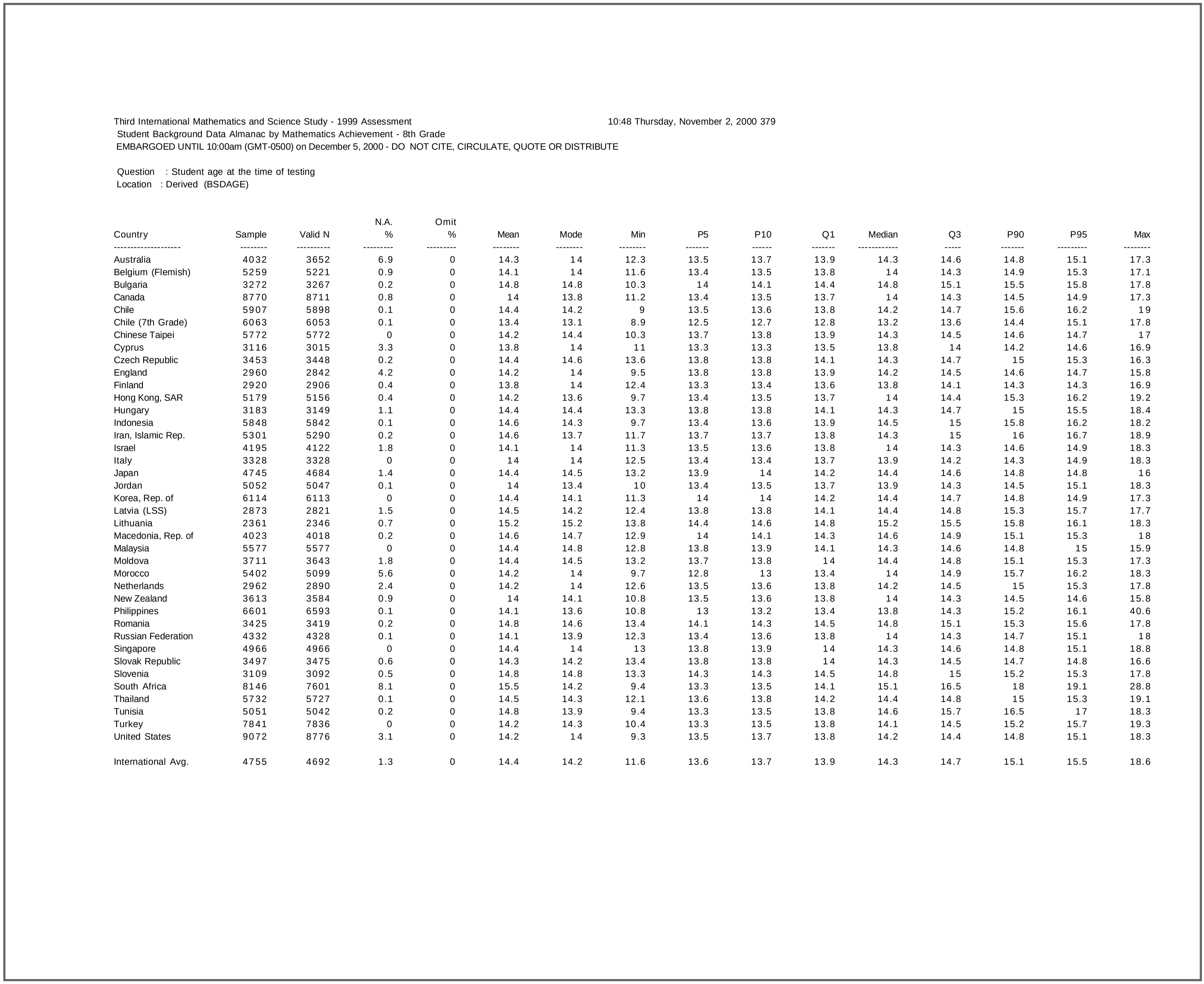
```
Third International Mathematics and Science Study - 1999 Assessment                          10:48 Thursday, November 2, 2000   4
Student Background Data Almanac by Mathematics Achievement - 8th Grade
EMBARGOED UNTIL 10:00am (GMT-0500) on December 5, 2000 - DO  NOT CITE, CIRCULATE, QUOTE OR DISTRIBUTE

Question   : Are you a boy or a girl?
Location   : SQ2-2 / SQ2S-2  (BSBGSEX)

                                                                NOT                                                  NOT
                                                             ADMINIS                                              ADMINIS
                                           1.GIRL     2.BOY    TERED     OMIT    1.GIRL     2.BOY     TERED      OMIT
Country                 Sample  Valid N         %         %        %        %      Mean      Mean      Mean      Mean
--------------------  -------- --------  -------- --------- --------- -------- --------- --------  --------- ---------
Australia                 4032     3940      50.7      49.3      2.3        0     523.4     526.7     527.1         .
Belgium (Flemish)         5259     5218      49.7      50.3      0.9        0     561.1     555.8     508.7     594.8
Bulgaria                  3272     3235      51.3      48.7      0.5      0.6     510.6       511     498.4     484.2
Canada                    8770     7558      50.4      49.6      0.9     12.6     531.4     535.2     513.6     514.8
Chile                     5907     5887      50.1      49.9      0.3      0.1     388.3     396.9     368.7     309.3
Chile (7th Grade)         6063     6034      48.3      51.7      0.3      0.2       355     363.4     328.1     323.8
Chinese Taipei            5772     5765      50.2      49.8        0      0.1     583.6     587.1     567.9     401.9
Cyprus                    3116     3096      49.1      50.9      0.7        0     478.6     475.6     379.3         .
Czech Republic            3453     3448      51.5      48.5      0.2        0     511.8     528.5     499.2         .
England                   2960     2841        49        51      4.1      0.1     487.2     507.1     474.9     358.4
Finland                   2920     2902      50.4      49.6      0.4      0.1     518.8     522.4     523.6     419.9
Hong Kong, SAR            5179     5098      49.4      50.6      0.4      1.1     583.7     582.3     492.1     536.3
Hungary                   3183     3166      50.4      49.6      0.4      0.2     528.6     534.7     534.3     495.4
Indonesia                 5848     5829      50.5      49.5      0.2      0.2     401.2     405.6     386.4     295.2
Iran, Islamic Rep.        5301     5301      40.6      59.4        0        0     408.1     431.8         .         .
Israel                    4195     4072      51.4      48.6      1.6      1.5     459.4     478.2     401.5     393.4
Italy                     3328     3328      51.1      48.9        0        0     474.9     484.2         .         .
Japan                     4745     4686      49.5      50.5      1.3      0.1     574.8     582.5       574     523.3
Jordan                    5052     5016        47        53      0.1      0.6     431.2     425.7     382.7     332.1
Korea, Rep. of            6114     6113      49.2      50.8        0        0     584.4     589.9       604         .
Latvia (LSS)              2873     2813      51.4      48.6      1.7      0.2     502.3     507.7     516.7       466
Lithuania                 2361     2339      51.7      48.3      0.7      0.3     480.1     483.3     505.7     397.5
Macedonia, Rep. of        4023     4000      49.4      50.6      0.2      0.4     446.9       447     304.1     421.4
Malaysia                  5577     5577      54.4      45.6        0        0     521.4     516.7         .         .
Moldova                   3711     3612      54.2      45.8        2      0.6     468.3       472     445.8     431.3
Morocco                   5402     5262      41.6      58.4      1.6        1     326.7     344.2     318.9     329.6
Netherlands               2962     2883      52.3      47.7      2.5      0.2     538.3     544.2     506.7     377.5
New Zealand               3613     3584      49.3      50.7      0.9        0     495.3     487.9     423.9         .
Philippines               6601     6555      53.4      46.6        0      0.6     352.5     337.2     279.3     278.2
Romania                   3425     3393      49.3      50.7      0.9      0.1     476.2       471     364.7     370.4
Russian Federation        4332     4329      52.1      47.9      0.1        0     525.7     526.4     509.9         .
Singapore                 4966     4964      48.5      51.5        0        0     603.3     605.5     551.9         .
Slovak Republic           3497     3471      50.4      49.6      0.8        0     531.4       536     571.6     519.2
Slovenia                  3109     3086        52        48      0.5      0.2     529.7     531.6       443     520.2
South  Africa             8146     8025      53.2      46.8      0.7      0.9     267.8     283.7     222.7     229.6
Thailand                  5732     5727      54.3      45.7      0.1        0     469.2     465.3     437.6         .
Tunisia                   5051     5020      50.9      49.1      0.4      0.1     435.9     460.7     418.7     429.1
Turkey                    7841     7834      42.1      57.9      0.1        0     427.7     429.3     321.3       452
United  States            9072     8797      50.2      49.8      2.2      0.3     499.2     507.1     436.9     489.7

International  Avg.       4755     4678        50        50      0.8      0.6     485.2     489.9     451.9     421.1
```

Exhibit 10.3 Example Data Almanac Display for a Numerical School Background Variable



Third International Mathematics and Science Study - 1999 Assessment 10:48 Thursday, November 2, 2000 379
Student Background Data Almanac by Mathematics Achievement - 8th Grade
EMBARGOED UNTIL 10:00am (GMT-0500) on December 5, 2000 - DO NOT CITE, CIRCULATE, QUOTE OR DISTRIBUTE

Question : Student age at the time of testing
Location : Derived (BSDAGE)

			N.A.	Omit											
Country	Sample	Valid N	%	%	Mean	Mode	Min	P5	P10	Q1	Median	Q3	P90	P95	Max
Australia	4032	3652	6.9	0	14.3	14	12.3	13.5	13.7	13.9	14.3	14.6	14.8	15.1	17.3
Belgium (Flemish)	5259	5221	0.9	0	14.1	14	11.6	13.4	13.5	13.8	14	14.3	14.9	15.3	17.1
Bulgaria	3272	3267	0.2	0	14.8	14.8	10.3	14	14.1	14.4	14.8	15.1	15.5	15.8	17.8
Canada	8770	8711	0.8	0	14	13.8	11.2	13.4	13.5	13.7	14	14.3	14.5	14.9	17.3
Chile	5907	5898	0.1	0	14.4	14.2	9	13.5	13.6	13.8	14.2	14.7	15.6	16.2	19
Chile (7th Grade)	6063	6053	0.1	0	13.4	13.1	8.9	12.5	12.7	12.8	13.2	13.6	14.4	15.1	17.8
Chinese Taipei	5772	5772	0	0	14.2	14.4	10.3	13.7	13.8	13.9	14.3	14.5	14.6	14.7	17
Cyprus	3116	3015	3.3	0	13.8	14	11	13.3	13.3	13.5	13.8	14	14.2	14.6	16.9
Czech Republic	3453	3448	0.2	0	14.4	14.6	13.6	13.8	13.8	14.1	14.3	14.7	15	15.3	16.3
England	2960	2842	4.2	0	14.2	14	9.5	13.8	13.8	13.9	14.2	14.5	14.6	14.7	15.8
Finland	2920	2906	0.4	0	13.8	14	12.4	13.3	13.4	13.6	13.8	14.1	14.3	14.3	16.9
Hong Kong, SAR	5179	5156	0.4	0	14.2	13.6	9.7	13.4	13.5	13.7	14	14.4	15.3	16.2	19.2
Hungary	3183	3149	1.1	0	14.4	14.4	13.3	13.8	13.8	14.1	14.3	14.7	15	15.5	18.4
Indonesia	5848	5842	0.1	0	14.6	14.3	9.7	13.4	13.6	13.9	14.5	15	15.8	16.2	18.2
Iran, Islamic Rep.	5301	5290	0.2	0	14.6	13.7	11.7	13.7	13.7	13.8	14.3	15	16	16.7	18.9
Israel	4195	4122	1.8	0	14.1	14	11.3	13.5	13.6	13.8	14	14.3	14.6	14.9	18.3
Italy	3328	3328	0	0	14	14	12.5	13.4	13.4	13.7	13.9	14.2	14.3	14.9	18.3
Japan	4745	4684	1.4	0	14.4	14.5	13.2	13.9	14	14.2	14.4	14.6	14.8	14.8	16
Jordan	5052	5047	0.1	0	14	13.4	10	13.4	13.5	13.7	13.9	14.3	14.5	15.1	18.3
Korea, Rep. of	6114	6113	0	0	14.4	14.1	11.3	14	14	14.2	14.4	14.7	14.8	14.9	17.3
Latvia (LSS)	2873	2821	1.5	0	14.5	14.2	12.4	13.8	13.8	14.1	14.4	14.8	15.3	15.7	17.7
Lithuania	2361	2346	0.7	0	15.2	15.2	13.8	14.4	14.6	14.8	15.2	15.5	15.8	16.1	18.3
Macedonia, Rep. of	4023	4018	0.2	0	14.6	14.7	12.9	14	14.1	14.3	14.6	14.9	15.1	15.3	18
Malaysia	5577	5577	0	0	14.4	14.8	12.8	13.8	13.9	14.1	14.3	14.6	14.8	15	15.9
Moldova	3711	3643	1.8	0	14.4	14.5	13.2	13.7	13.8	14	14.4	14.8	15.1	15.3	17.3
Morocco	5402	5099	5.6	0	14.2	14	9.7	12.8	13	13.4	14	14.9	15.7	16.2	18.3
Netherlands	2962	2890	2.4	0	14.2	14	12.6	13.5	13.6	13.8	14.2	14.5	15	15.3	17.8
New Zealand	3613	3584	0.9	0	14	14.1	10.8	13.5	13.6	13.8	14	14.3	14.5	14.6	15.8
Philippines	6601	6593	0.1	0	14.1	13.6	10.8	13	13.2	13.4	13.8	14.3	15.2	16.1	40.6
Romania	3425	3419	0.2	0	14.8	14.6	13.4	14.1	14.3	14.5	14.8	15.1	15.3	15.6	17.8
Russian Federation	4332	4328	0.1	0	14.1	13.9	12.3	13.4	13.6	13.8	14	14.3	14.7	15.1	18
Singapore	4966	4966	0	0	14.4	14	13	13.8	13.9	14	14.3	14.6	14.8	15.1	18.8
Slovak Republic	3497	3475	0.6	0	14.3	14.2	13.4	13.8	13.8	14	14.3	14.5	14.7	14.8	16.6
Slovenia	3109	3092	0.5	0	14.8	14.8	13.3	14.3	14.3	14.5	14.8	15	15.2	15.3	17.8
South Africa	8146	7601	8.1	0	15.5	14.2	9.4	13.3	13.5	14.1	15.1	16.5	18	19.1	28.8
Thailand	5732	5727	0.1	0	14.5	14.3	12.1	13.6	13.8	14.2	14.4	14.8	15	15.3	19.1
Tunisia	5051	5042	0.2	0	14.8	13.9	9.4	13.3	13.5	13.8	14.6	15.7	16.5	17	18.3
Turkey	7841	7836	0	0	14.2	14.3	10.4	13.3	13.5	13.8	14.1	14.5	15.2	15.7	19.3
United States	9072	8776	3.1	0	14.2	14	9.3	13.5	13.7	13.8	14.2	14.4	14.8	15.1	18.3
International Avg.	4755	4692	1.3	0	14.4	14.2	11.6	13.6	13.7	13.9	14.3	14.7	15.1	15.5	18.6

10.5.2 Versions of the National Data Files

Building the international database was an iterative process. The IEA Data Processing Center provided NRCs with a new version of their country's data files whenever a major step in data processing was completed. This also guaranteed that the NRCs had a chance to review their data and run their own checks to validate the data files.

Three versions of the data files were sent out to the countries before the TIMSS international database was made available. Each country received its own data only. The first version was sent to the NRC as soon as that country's data had been cleaned. These files contained nationally standardized achievement scores calculated by the Data Processing Center using a Rasch-based scaling method. Documentation, with a list of the cleaning checks and all corrections made in the data, was included to enable the NRC to review the cleaning process. Univariate statistics for the background data and item statistics for the achievement data were also provided for statistical review. A second version of the data files was sent to the NRCs when the weights and the international achievement scores were available and had been merged with the files. A third version was sent together with the data almanacs after final updates had been made, to enable the NRCs to validate the results presented in the first international reports.

10.5.3 Reports

Several reports were produced during data processing at the IEA Data Processing Center to inform and assist the NRCs, the TIMSS International Study Center, and other institutions involved in TIMSS 1999. The NRCs were provided with diagnostic reports and univariate statistics to help them check their data. The TIMSS International Study Center and ETS were provided with international item statistics. The International Study Center also received international coding reliability statistics and international univariate statistics. A report was made to the International Study Center and the TIMSS 1999 Project Management Committee about the status of each country's data, any problems encountered in the data cleaning, and general statistics about the number of observations per file and preliminary student response rates.

10.6 Computer Software

dBase was used as the standard database program for handling the incoming data. Tools for precleaning and programs such as LINKCHCK (described earlier) and MANCORR and CLEAN (described below) were developed for manipulating the data. Statistical analyses (e.g., univariate statistics) for data cleaning and review were carried out with SAS. The final data sets were also created using SAS. For item statistics, the Data Processing Center used the QUEST software (Adams and Khoo, 1993).

The main programs that were developed by the Data Processing Center and used for TIMSS 1999 are described below. Most of the programs that were written for country-specific cleaning needs are not listed. The programming resources in the main cleaning process were spent largely in developing this set of programs.

10.6.1 MANCORR

The most time-consuming and error-prone part of data cleaning is the direct or 'manual' editing of errors uncovered by the review process. Based on the Data Processing Center's experience in the IEA Reading Literacy Study, TIMSS 1995, and the pilot phases of TIMSS 1999, the data-editing program MANCORR was developed. It is easy to use and generates automatic reports of all data manipulation. Its main advantage compared with other editors is that all changes in the data are documented in a log database, from which reports can be generated. As updated data were received from countries, the time-intensive manual changes could be automatically repeated. An 'Undo' function allowed the restoration of original values that had been modified with the MANCORR program. The report on which changes were made in the data, by whom, and when was important for internal quality control and review. The MANCORR program was designed using CLIPPER in order to manipulate DATAENTRYMANAGER files.

10.6.2 CLEAN

The main software instrument for data cleaning in TIMSS 1999 was the diagnostic program CLEAN. This program was derived from earlier versions used in the IEA Reading Literacy Study and TIMSS 1995. It was used to check all the TIMSS 1999 files individually, the linkages across files, and all between-file comparisons. An important feature of the program is that it could be used on a data file as often as necessary. It could first be used to make automatic corrections, and subsequently for creating a report only, without making corrections. Thus it was possible to run a check on the files at all stages of work until the end, when the file format was changed to the SAS format. This meant that the program was used not only for initial checks but also to check the work done at the Data Processing Center.

A feature of the TIMSS 1999 data cleaning tools is that all deviations are reported to a database, so that reports can be generated by type of problem or by record. Reports previously generated by the program could be compared automatically with newer reports to see which problems had been solved, and even more important, whether additional deviations were introduced during manual correction. The databases were used to generate the final reports to be sent to the countries. These reports showed which deviations were initially in the data, which were solved automatically, which were solved manually, and which remained unchanged.

10.6.3 Programs Creating Meta Databases

Using SAS, several programs were developed by the Data Processing Center for reviewing and analyzing both the background data and the test items. For the background data, a meta database containing information provided by the initial analysis and by the international codebook was created. Another meta database containing the relevant item parameters was created for the achievement test items. Later, all statistical checks and reports used these databases instead of running the statistics over all data sets again and again. If the data for one country were changed then statistics had to be recalculated for that country only. This reduced the computing time for certain procedures from hours to a few minutes. Both databases are the base sources of several reports produced at both the national and international levels (e.g., for the univariate and item analysis reports).

10.6.4 Export Programs

As mentioned above, SAS was the main program for analyzing the data. Using SAS, export programs were developed and tested to create output data sets for data distribution that are readable by either SAS or SPSS.

10.7 Summary

The structures and procedures designed for processing the TIMSS 1999 data were found to be very successful. In planning for the TIMSS data processing, the major problems were anticipated and provision for dealing with them was incorporated into the data processing system. The IEA Data Processing Center was closely involved in the planning phase of the study. The study thus benefited from the Center's knowledge and experience in data processing. TIMSS 1999 also benefited from the experience gained with TIMSS in 1995. Procedures and practices developed in response to problems encountered in the earlier study were refined and made even more effective in 1999. In particular, since the work of checking and cleaning the TIMSS database required a vast amount of interaction between NRCs in each country, the IEA Data Processing Center in Hamburg, the TIMSS International Study Center in Boston, Statistics Canada in Ottawa, and Educational Testing Service in Princeton, improvements in communications and data transmission greatly facilitated the entire enterprise.

References

Adams, R.J., & Khoo, S. (1993). *Quest: The interactive test analysis system.* Melbourne: Australian Council for Educational Research.

Gonzalez, E.J., & Smith, T.A., Eds. (1997). *User Guide for the TIMSS international database: Primary and middle school years – 1995 assessment.* Chestnut Hill, MA: Boston College.

TIMSS (1998). *Manual for Entering the TIMSS-R Data* (Doc. Ref. No.: 98-0028). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.