# Resolving Inconsistencies in the TIMSS and PIRLS 2011 Data

Oliver Neuschmidt

The number of inconsistent or implausible responses in the data files varied from country to country, but no country's data were completely free of inconsistent responses. Each problem was recorded in a database along with a description of the problem. Issues that could not be resolved using systematic cleaning rules were reported to the NRC so that original data-collection instruments and tracking forms could be cross-checked. Where the national center could not solve the problems by inspecting the instruments and forms, a set of final cleaning rules were applied.

As outlined in **Creating and Checking the TIMSS and PIRLS 2011 Databases**, the cleaning process involved the adherence to a number of standardized cleaning rules. The cleaning rules were implemented to ensure that the data from countries and benchmarking entities were processed in a uniform manner.

Listed below are some of the common inconsistencies encountered during the data cleaning process as well as information on how they were resolved:

- **Filter questions**
  Filter questions are preliminary questions that direct respondents away from questions that do not apply to them (for example: "Do you have a computer at home? If no, skip to next section"). In every country, a small minority of respondents answered questions in sections that were logically implausible. Inconsistencies between the coding of filter questions and their dependent questions usually were treated automatically by the cleaning programs. If the filter variable contained a valid value and its dependent questions were correctly skipped, dependent variables were coded as "logically not applicable". If a response to a filter question was equivalent to "No", indicating that the dependent questions were not applicable, but the dependent questions contained valid answers then the filter question was recoded to "Yes".

- **Split Variable Checks**
  Split variable checks were applied to "Yes/No" lists and percentage list questions where the responses were coded into several variables. For example, question 5 in the TIMSS and PIRLS Grade 4 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 eleven variables, each one was coded as "Yes" if the corresponding possession was checked or it was coded as "No" if left unchecked. Occasionally, students checked the "Yes" boxes but left the "No" boxes unchecked or missing. Since in these cases it could be assumed that the unmarked boxes actually meant "No", the corresponding variables were imputed accordingly. The individual responses to percentage questions were summed up and, if they fell outside the range of 90 to 110, they were coded to "omitted".

♦ **Implausible Numeric Value**
Variables with implausible numeric values were set to "omitted". For example, question 6A in the TIMSS and PIRLS Grade 4 School Questionnaires asked about the number of instructional days per year. Values exceeding 250 and values below 150 were coded to "omitted".

♦ **Mis-punched Missing Variables**
A focus of the data cleaning process was to identify possible mis-punched missing variables as these might bias the analysis results. The highest valid value to be entered for percentage questions for example is 100 and consequently any missing codes would need to be entered with a length of 3 digits (999 or 998). However, in some cases data entry staff entered two digit codes, such as '99' or '98' counting as a valid value within the defined variable range. Those possibly mis-punched missing codes were identified and wherever necessary corrected. Where problems could not be solved in collaboration with the country, the 95% percentile of the value distribution was calculated by country and possibly mis-punched missing values exceeding this threshold were recoded to 'omitted'.

♦ **Inconsistent Response Patterns**
Certain questions were verified against one another to identify and resolve inconsistent response patterns. For example, question 1A in the TIMSS and PIRLS School Questionnaires asked for the total school enrollment in all grades, while question 2 asked for the enrollment in the target grade only. Clearly, the number given for question 2 should not exceed the number given for question 1A.