Part

II

THE DATA COLLECTION

II.1 Introduction

Data collection consists of several general tasks:

- Creating data entry and validation programmes and protocols;
- Formatting countries' data into a consistent format;
- Preparing validity reports and statistical summaries of items to ensure accuracy;
- Data cleaning; and
- Merging countries' data files to create the OECD international database.

Each task was crucial to ensuring the quality and accuracy of the data file for the ISUSS.

This section describes the data entry task undertaken by each National Project Manager (NPM), the data checking and database creation that was implemented by the OECD international co-ordinator, and the steps taken to ensure the quality and accuracy of the international database. It discusses the respective responsibilities for each step in the process of creating the international database; the flow of the data files between those responsible for data processing; the structure of the data files submitted by each country for processing and the resulting files that represent the international database.

Data processing for ISUSS was carried out as a joint activity, with close cooperation between the NPM in each of the participating countries and the OECD international co-ordinator.

The change in schedules for data processing arising from the withdrawal of the original contractor were discussed and agreed at the first meeting of the NPMs in June 2001. There were eventual deviations from the negotiated schedules, but the system allowed all data processing to be done by the international co-ordinator. Moreover, the system proved to have advantages later in the iterative cleaning and editing processes because late data entries and edit changes could be added to the database.

II.2 Data flow

The data collected in the ISUSS survey were entered into data files using a common international format, as described in the ISUSS Codebook (Annex 4). The data files were then submitted to the OECD international co-ordinator for cleaning and verification. The major responsibilities of the OECD were to check that the data files received matched the international standard and to request modifications as necessary to bring them into compliance. With the exception of one country that lacked this capability and where OECD needed to reformat the data set, all countries carried out the reformatting to codebook standards.

Once the data sets were formatted, OECD then applied standard cleaning rules to the data to verify their consistency and accuracy. A software 'cleaning programme' has been developed to produce a validity report indicating any inconsistencies. The cleaning programme was iterative, so that NPMs could edit and re-run the validation programme to ensure the accuracy and internal consistency of their data.

Finally, when the data appeared to be clean and the validity report indicated no serious inconsistencies or format problems, OECD created statistical summaries for each country using their data. These statistical

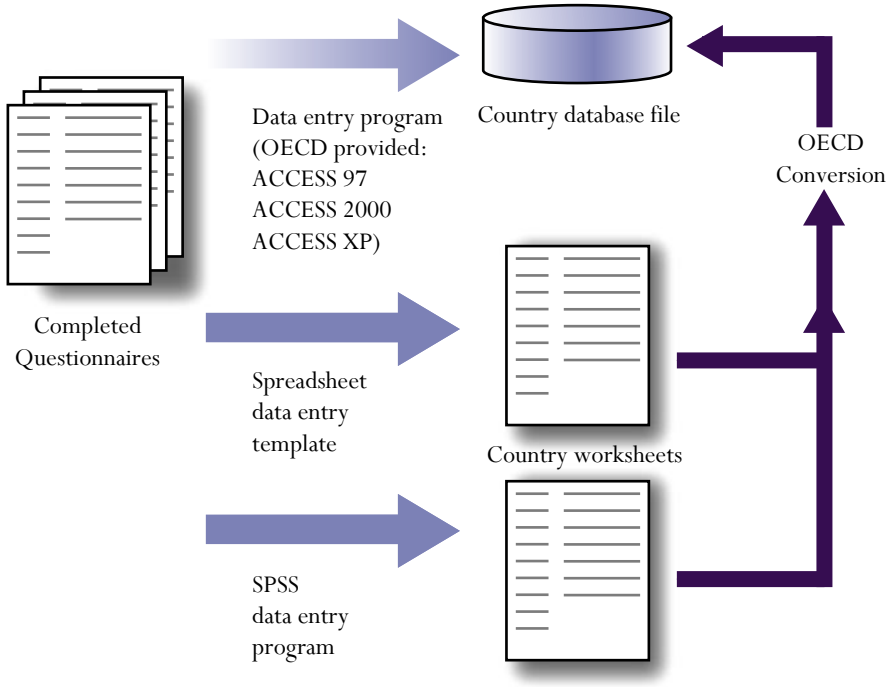
summaries (ranges, means, missing values, etc.) gave a general view of data patterns item by item, which provided yet another view of the data to expose any remaining problems.

Once countries had verified and approved their data and OECD had determined that it was consistent with the international file format, the data were merged into an international database. During the creation of the international database, country codes were added and a final edit check was run for internal consistency. A final summary file of basic item statistics was produced for each country and sent to them for their review. This review, together with data provided by the NPM tracking forms, was used to determine the acceptability of each country's data for the ISUSS international database.

II.3 Data entry by the national project managers

The diagram shown in Figure II.1 illustrates the various programmes and format conversions which were required from data entry to a final country database file.

Figure II.1
Data entry process



Each country's NPM was responsible for translating and transcribing the information from the national questionnaires into a format based on the generic ISUSS questionnaire and for creating a computer data file. There was some freedom to choose the method of data entry, as long as the format was consistent with the ISUSS codebook. About half of the countries chose to use data-entry software provided by OECD, less than half used SPSS to enter data, many of the remaining countries arranged to use spreadsheets as entry forms, and one country used SAS.

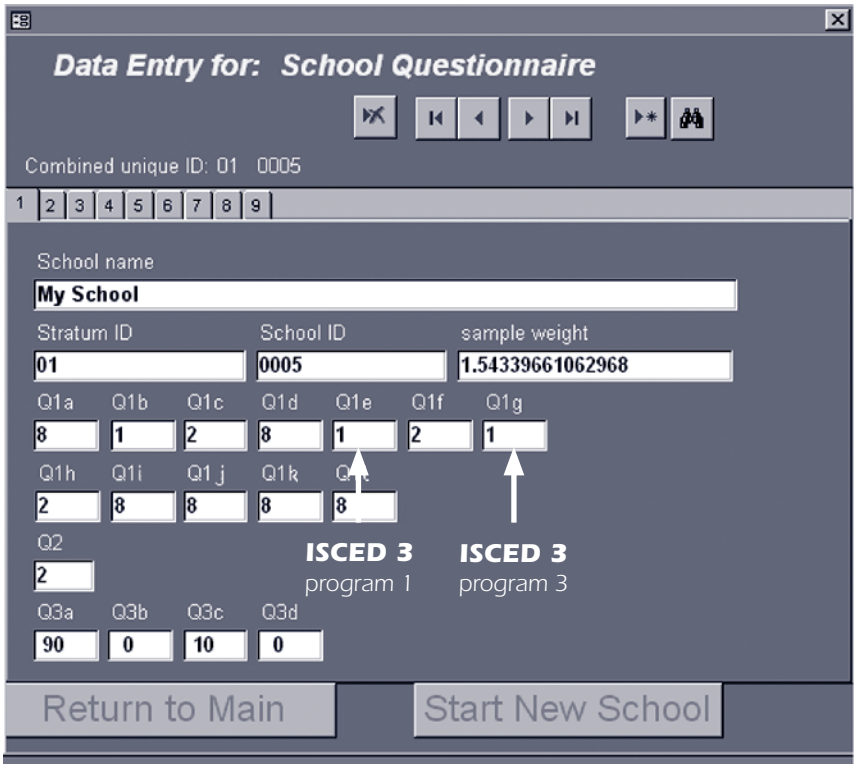
The Codebook (Annex 4) for the questionnaire was a key for assuring consistent and accurate data for ISUSS. There were two parts to the codebook, the school section which was to be completed for each school, and the programme section(s) to be completed for each upper secondary programme in the school. This structure was inherently hierarchical, with identification numbers to make sure that programmes were linked to the right school. This arrangement meant that there were data files for schools and for programmes within schools, linked as a hierarchy to reflect the arrangement of programmes within schools. The codebook contained the essential information for providing this link.

OECD developed data entry software to facilitate NPMs' work and to ensure compliance with the codebook requirements. There were two alternatives for those countries that had not already decided on a system of data entry: a spreadsheet that incorporated value screens to make data entry consistent with codebook requirements; and a data entry programme based on ACCESS, a database language system. Because countries had different versions of ACCESS, the database system was developed in three versions: ACCESS 97, ACCESS 2000 and ACCESS XP.

Because of the hierarchical nature of ISUSS data, with a variable number of programmes within each school, OECD recommended the use of the ACCESS database system. The alternative spreadsheet data entry system did not automatically ensure a correct link between a school and its programmes, which meant the possibility of error. To facilitate data entry, the ACCESS screens were designed to match the school and programme sections of the questionnaire. Not only did this make data entry easier, but it preserved the link between a school and its programmes.

Figure II.2

Opening data entry screen for School Questionnaire Data



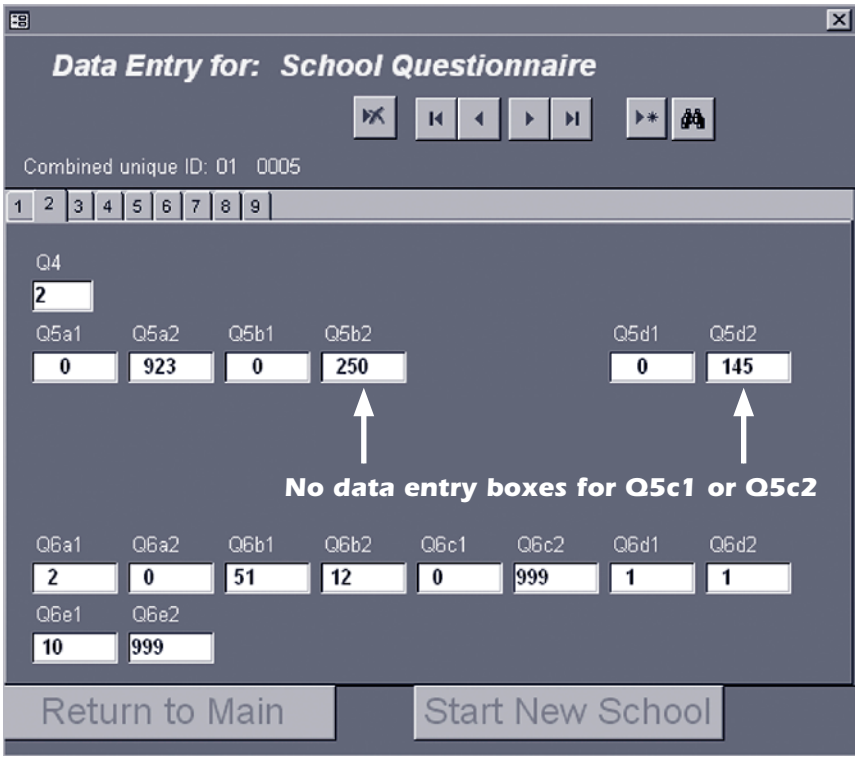
For example, the data entry screen for the first questionnaire item, shown in Figure II.2, illustrates a pattern of educational programmes from primary education through higher education. The responses indicate that the school provides two upper secondary programmes: the first and third in the country's list of programmes (indicated by '1' in Q1e and Q1g). The screen limits the number of data slots available to just those appropriate to this country: two upper secondary programmes.

Figure II.3 shows the second screen of the school data entry programme. This screen accepts responses to questionnaire item 5 about enrolments in the school for upper secondary programmes. The example shows that the inappropriate data entry boxes for item 5 have been blanked out for this school, leaving only the first and third upper secondary programmes available for data entry. This illustrates one of the advantages of using the OECD supplied software to ensure data accuracy and consistency – data could not be entered in an inappropriate box.

During the course of data entry, there were several changes made to the data entry programmes, in spite of all the pre-planning and pilot testing that had been done. Both the enrolment sizes in some upper secondary programmes and the numbers of computers available to students proved larger than anticipated. The data entry software needed to be modified to hold the larger numbers. It was a simple matter to make changes in the value screens and the codebook, but it turned out to be somewhat more difficult to change the missing values that had already been entered by those countries which had completed some of their data entry.

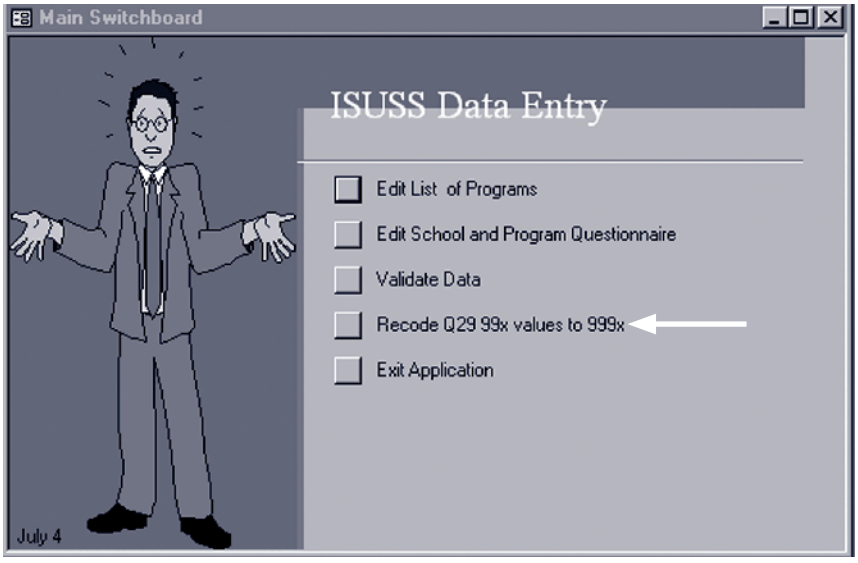
Figure II.3

Second data entry screen for School Questionnaire Data



To illustrate this problem, suppose that enrolment data had been entered before the change was made from a 4-digit to a 5-digit field. Before the change in value screens, only 4-digit data could be entered, with the maximum enrolments limited to 5 000 and special data codes set to '9997', '9998' and '9999'. After the modification, the maximum enrolments was 15 000 with special data codes set to '99997', '99998' and '99999'. Although it was extremely unlikely that one of the 4-digit special data codes would correspond to actual enrolments, this possibility still required checking to avoid converting an actual enrolments figure to a missing value. Figure II.4 illustrates one attempt to correct for this problem, after confirming with the countries that the conversion would not create inappropriate missing values.

Figure II.4
Data cleaning module opening screen

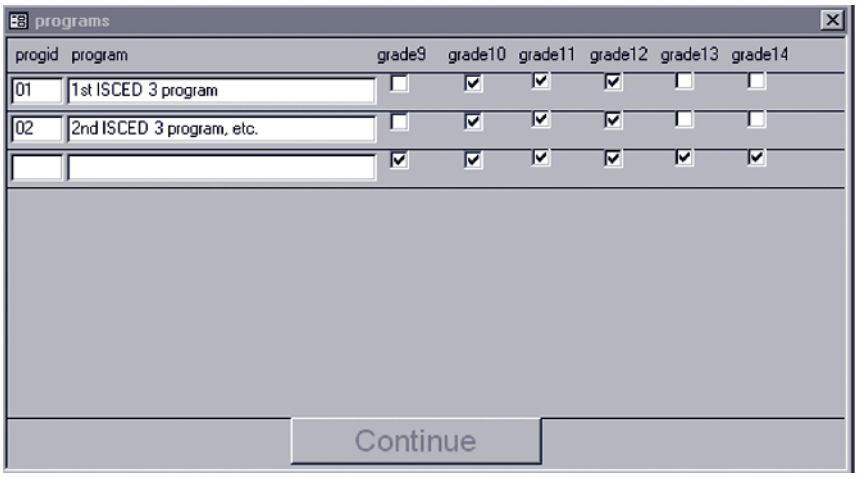


Another change was made to the codebook and data entry programmes during data entry to include information on grade equivalencies for specific upper secondary programmes. It was discovered during routine data checks that several upper secondary programmes were directed to older or younger students, although the ISCED level was the same. To ignore this difference would complicate subsequent analyses of programme data. Therefore, although this information was neither a questionnaire item nor listed in the codebook, a data entry screen was added to the ACCESS programmes to gather additional data on the cumulative years of schooling typical for each year of each programme. Since the data could be accurately completed by the NPM from administrative records, it did not significantly add to the data burden on respondents, so the decision was made to add the information to each country's data file.

For countries not using the ACCESS programme for data entry, the programme by grade screen was included in the data cleaning module so that it could be added when the country made validation edits. The information request is shown in Figure II.5. The example shows a country where two ISCED 3 programmes are available. In each of these, students enter the programme after nine years of schooling in ISCED 1 and ISCED 2 programmes; ISCED 3 programmes are three-year programmes. The examples in the subsequent figures refer to this case.

Figure II.5

Additional data entry form on grade level



The most difficult part of the questionnaire for respondents and for data entry was Part III, the upper secondary programme data. Figure II.6 shows the first screen for data entry of Part III of the questionnaire. The screen ensures that the correct upper secondary programme was selected by presenting its name in an un-editable field.

Figure II.6

First data entry screen for School Programme Data

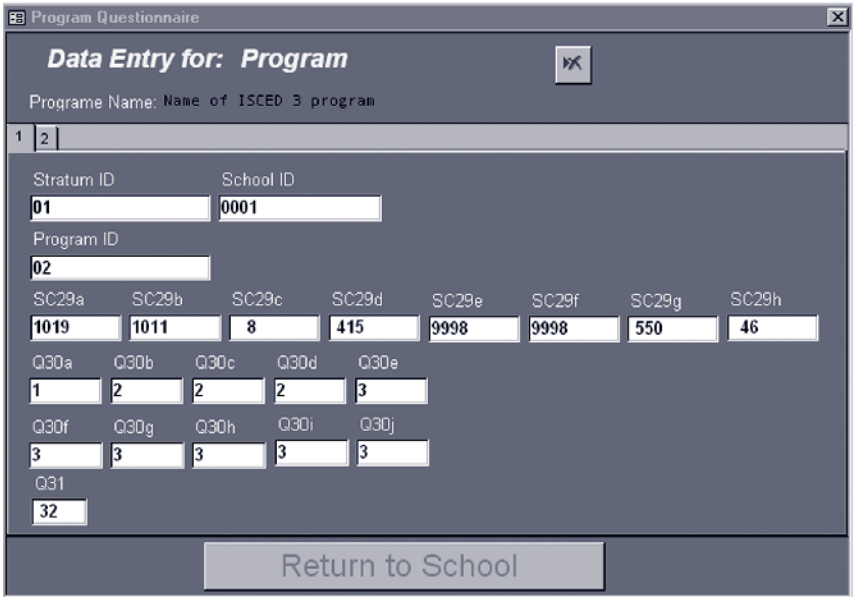
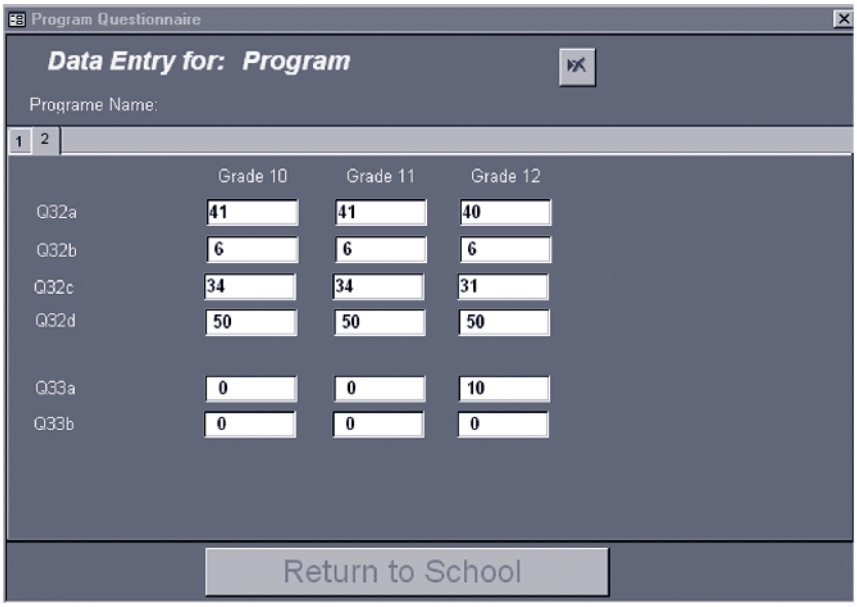


Figure II.7 shows the second page of the data entry screen for programme information. Here the data cells for grades 9, 13, and 14 are missing, since the programme by grade data as shown in Figure II.5 indicates that this programme cannot have these grades enrolled. These restrictions on data entry significantly reduced data entry errors.

Figure II.7
Second data entry screen for School Programme Data



II.4 Data cleaning

As already noted, not all countries used the data entry programmes provided by OECD. Accordingly, the data entry files submitted by each country differed, depending on the programme used. About half of the countries submitted ACCESS databases, although in three versions. Unfortunately, there was little compatibility between the versions of this programme, so a single version, ACCESS 97, was selected as the ISUSS standard. This provided the needed consistency, although it meant time-consuming data conversions to fit the software requirements of the countries.

For countries using the spreadsheet data entry method, the conversion to an ACCESS 97 file was straightforward. However, some data manipulation was often required because the spreadsheet data entry created two sheets: one for school-level data; and one for upper secondary programme-level data. There was no automatic linkage that ensured an accurate match between the school and its upper secondary programmes, so there was the possibility of error both in data entry and in data conversion. Because of its hierarchical structure, however, the ACCESS database automatically ensured that any anomalies were discovered and could be fixed.

In order to check for the possibility of errors from data entry or from inconsistent responses, the ACCESS data entry programme was enhanced to provide extensive checking of the data. There were consistency and range checks to verify the accuracy and completeness of the data. In addition, the data cleaning

module included editing capabilities and report generators to produce data validity reports and summary information to be used to pinpoint possible data errors.

The kinds of checks in the data cleaning module were generally of the following sort:

- Codebook format checks: Conformity to coding of variable names and locations;
- Range checks:
  - Allowable values for each item with a Yes/No response; and
  - Reasonable ranges for each item with numeric data response.
- Consistency checks:
  - Numbers of programmes match positive enrolment figures;
  - Percentages rounded to 100 per cent; and
  - Numbers that are parts of a whole are smaller than the whole.
- Appropriate codes for illegal or missing data.

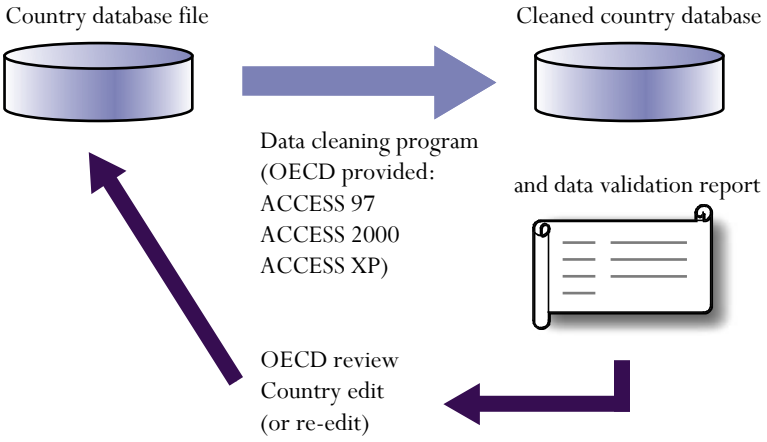
The ISUSS data cleaning process was designed to identify, document and, where possible, help countries correct deviations from the international file structure of the codebook. This involved developing reports to identify data entry errors and systematic deviations from the international data formats, identifying problems in linking observations across school and programme files, and noting any inconsistencies within and across observations. The objective was to ensure that the ISUSS data adhered to international formats and accurately and consistently reflected the information collected within each country.

Data cleaning involved several steps and was an iterative process, as shown in Figure II.8: some of the steps needed to be 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 in an ACCESS database. As a second step, all problems with identification variables, linkage across school and programme files, codes used for different categories, and missing value categories were detected and reported.

At the last stage of data cleaning, a series of statistical summary reports were generated for each country. The reports contained listings of codes used for each item in both school and programme sections of the questionnaire and pointed to outliers, ranges, missing value patterns and other codes that indicated problematic data (not administered, ambiguous or multiple entry codes). The reports were sent to each participating country using only the data from that country. NPMs were asked to review the data and make changes in the cleaning programme as necessary. In many cases the NPM needed to return to the original questionnaires to resolve questions and update or edit the cleaning file.

In almost all cases, corrections to the data were done by the country using the data cleaning module. In some cases, countries asked for help in making changes to individual school or programme data entries. In all cases where OECD made any change to a country's data, a data validation report was sent to the country along with an updated cleaning programme with the newly edited data. NPMs were asked to verify that the changes had been made in accordance with their instructions. Only in one country was the

Figure II.8
Iterative data cleaning process



Optional: Spreadsheet conversions for those countries without ACCESS 97, ACCESS 2000 or ACCESS XP. OECD ran data validation and returned the data validation report to countries for edits on the spreadsheet data input form.

NPM not able to make the necessary changes and OECD needed to transform and change data to conform to codebook standards. Even in this case the country reviewed and approved the final iteration of the data validation report.

All changes made to countries' databases were documented. Copies of databases were created that give a chronological file, should there ever be a need to return to an earlier version or discover where an error was made. Thus it is possible to reconstruct each original database received from a country.

II.5 Standardisation of the international database

The final step in creating the ISUSS international database was to merge all countries' individual data files into a single file. A final validity check was run to ensure that the data were consistent with the required file format and that value ranges were legal and logical. Since the individual country files did not include country names, the merged data also added abbreviated country alphabetic and numeric codes.

It was not clear at the outset whether it would be more useful to have the final format of the ISUSS international data as a database, such as ACCESS, or as a statistical package, such as SPSS. It was therefore decided to produce both formats. However, experience has shown that version control can be a difficult problem in keeping track of final data files. Therefore, the database and statistical formats were created together to prevent different versions in each format.

The ACCESS database was considered the master file and a specific programme was created to make the SPSS conversion from the ACCESS database. During the data cleaning phase, all changes in data were made on the ACCESS database and transferred to the SPSS file by using the conversion programme. All analyses have been done using the SPSS database.