

Standard Operating Procedures

Subject: Data Entry

Date: 15Jun2006

Purpose: The purpose of this Standard Operating Procedure (SOP) is to define the minimum standards for data entry. Research projects managed by personnel at FPG Data Management and Analysis Center (DMAC) process raw data in a wide variety of ways. This document will focus on the method that is most often used, the process of converting raw data on paper forms to SAS data sets. Processing raw data in other mediums will be briefly described.

Scope: This procedure applies to all projects managed and maintained at the FPG Data Management and Analysis Center.

Responsibility: Setting up a data entry system is the responsibility of the programmer assigned to the project. If more than one programmer is assigned to a project, the senior programmer on a project is responsible for establishing the style and pattern of data entry.

Independent Double Data Entry with a Third Party Referee: This method of data entry is the most common in the Data Management and Analysis Center. Data that has been collected on paper forms are given in batches to the programmer assigned to the project.

- Each item (form, survey, etc.) must have a unique identification number, or a unique combination of several fields.
- SAS data entry screens are designed, and ranges set, required fields indicated, etc, where appropriate.
- Data is entered by two individuals, each working independently of the other, into two separate data sets.
- Data entry personnel should be instructed to “key what they see”. Deviations from this rule should be clearly marked during a pre-entry review process.
- The two data sets are compared by a third person, who identifies and reconciles discrepancies.
- The resolved data set should be permanently saved as a separate data set, as well as added to the data stream for that data type.

Other Types of Data Entry:

From time to time raw data is received in a variety of other forms. These forms include (but are not limited to) flat ASCII files, Excel spreadsheets, Access databases, Word files, machine generated files and Internet data entry files. It is the responsibility of the programmers assigned to the project to determine the most appropriate way to convert these forms of raw data into SAS datasets.

Incorporating New Data into the Main Data Stream:

After a batch of data has been keyed and reconciled, another set of programs adds individual batches to the main data stream for each instrument.

- A permanent label should be assigned to each SAS variable.
- Dates should be stored as SAS “date” variables, with permanent SAS formats to display the four-digit year.
- Error detection code should check all variables for valid values and consistency with other variables as appropriate for the form in question.
- A variable should be created that will identify the batch in which each observation entered the datastream.
- After the merge or set of new data into the permanent data stream, any new data should be printed as a subset of the larger dataset.
- A randomly selected subset of any new data may be printed for manual verification.
- ID variables and other key variables such as gender and date of birth should be checked against a master dataset.
- Within each program subdirectory, an index file should list and briefly describe each program. A master list of the forms for a project and the status of the datasets for each should also be maintained.
- A master list of the forms for a project should be maintained.

Verification of Data:

- New data should be printed for checking, *after* the reconciled (or new) data is added to existing data set.
- A thorough check of some minimum number of records should be made by DMAC personnel and/or by project personnel. The actual number of records to be checked is agreed upon by those involved. This is done to assure that data values have been assigned to the correct variables, and that no unexpected incompatibilities (i.e. length, type, etc.) occurred when new data were merged or set with existing data sets.
- Calculated variables should be carefully checked by DMAC personnel to assure that the calculations are correct, and an additional check by project personnel is recommended to assure that the instructions have been correctly implemented.

Definitions:

- Reconciliation: The process of electronically comparing the two independently keyed versions of a batch of data and making a decision about any discrepancies. This is done by a third person, and the decisions of that person are incorporated into a third dataset that is then added to the main data stream for that particular instrument.