

Standard Operating Procedures

Subject: Database Design and Management

Date: 15Jun2006

Purpose: The purpose of this Standard Operating Procedure (SOP) is to define the minimum standards of a database design and management system. The purpose of a DBMS is to ultimately assure high quality data for statistical analysis. This SOP briefly covers topics that are covered in more detail in individual SOPs.

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

Responsibility: Database design 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 database design. Each programmer is responsible for maintaining consistency of database design throughout a project. It is the responsibility of the senior (or supervising) programmer to ascertain that adequate and consistent design is maintained.

Components of Database Design: The FPG Data Management and Analysis Center recommends that the DBMS design include all of the following components. Whenever funding permits, these components should be present in a DBMS design.

- Review of proposed work and forms
 - Whenever possible, the programmer will review the proposed work and data collection instruments prior to data collection in order to assure that the data will be in a format suitable for ease of interpretation as well as efficient and accurate data entry.
- Development of ID system
 - It is essential that the programmer help a project establish a consistent ID system.
 - Each piece of data must have a unique identifier or set of identifiers.
 - These identifiers must assure that each piece of data can be appropriately matched with other related pieces of data within the same project and across projects when the projects are inter-related.
- Inventory, tracking, controlling files/records
 - An inventory list may be kept of the forms sent to be processed, received by DMAC, processed and returned to the project after processing.
 - A record may be kept of the status of the project datasets such that an overview of the readiness of the data is available for review.
 - A computerized inventory of the forms processed for each subject may be made which enables a check for completeness across measures.
 - A controlling computerized file may be kept against which all pieces of data may be checked for consistent ids, birth dates, gender, and other important information that needs to be consistent across measures.

- Directory structure and naming conventions
 - All project data should be stored on network drives with a root node name identifying the project that is stored under that root.
 - The directory structure should reflect the organizational structure of the data collection instruments.
 - Individual datasets should be named to reflect the instrument they represent, and each version should have a unique, sequential name.
 - Each directory should have an index explaining what is stored therein.
- Data entry
 - Data arriving on paper forms should be independently double keyed, electronically compared and reconciled by a third person.
 - Data arriving in other formats should be processed using methods appropriate to the format in which it arrived.
- Creating/adding to a database
 - Unique filenames should be used to allow reconstruction of the database if ever necessary.
 - Scoring plans should be determined as early as possible in the development of the database. Written plans for scoring should be supplied by project staff and/or statisticians. Once in place, scoring is re-run every time the database is updated in order to assure that any corrections are automatically incorporated in scoring. Scoring should be confirmed by at least a second person.
 - Error reports should be written to bring attention of project staff to any invalid or illogical responses. Project staff should respond to these reports in writing, and all corrections should be made until the data are clean.
 - An electronic comparison should be run comparing each version of a database to the previous version in order to check that all intended changes have indeed been made and that no unintended changes have been made.
 - A careful check of number of observations and number of variables should be made to assure that these are always as intended.
 - All logs and outputs should be carefully examined for possible errors.
- Data and scoring verification
 - Programming staff should verify at least a portion of all new data.
 - Additionally, the initial version of a database and associated scores should be completely verified by project staff, and subsequent versions should be verified by project staff at least 10%.
- Documentation
 - A codebook and annotated form should be created to document each data stream.
 - Programs should include comments that explain what the program is doing.
- Supervision
 - New employees should be trained through use of the written DMAC training manual, by example, by written and oral direction from supervisors/upper level programmers.
 - Supervisors and/or upper level programmers should thoroughly review all code written by new employees/junior programmers until satisfied with the consistent accuracy of the code written.

- Once satisfied with the consistent accuracy of code written, supervisors/upper level programmers should still review a selection of programs written for accuracy.
- Project staff should verify a portion of each dataset to assure accuracy of data entry and any calculated variables.
- Confidentiality and security
 - Employees will sign a confidentiality pledge
 - The computer support personnel will establish network security procedures in accordance with common practice
- Back-up and Archival
 - All network drives should be backed up in accordance with the established schedule.
 - When a current project is completed, its data should be archived to CDs with copies kept at FPG, stored off-site, and given to the PI.
 - When funding permits, a set of summary notebooks and CDs should be created to facilitate any future use of the data.

Definitions:

- Database management design system (DBMS): A collection of manual procedures, data files, computer programs and documentation, all of which are used to process data through a sequence of steps, ultimately producing high quality data files ready for statistical analysis.
- Data Stream: A set of files documenting the evolution of the data.