DATA PROCESSING AND DATA BASE CONSIDERATIONS FOR THE NEW NLS COHORT

Karin Steinbrenner, National Opinion Research Center

Introduction

The new cohort of the National Longitudinal Studies of Labor Force Behavior is a longitudinal survey in which a sample of 12,000 youth will be interviewed once a year over a period of six years.

The initial sample will be derived from a screener survey in which approximately 90,000 households will be enumerated with an average of 3.5 individuals in each family unit. The screening process will start in September and continue through December.

From the screener data (which in itself is a hierarchical file: family unit data, individual data, address information for some individuals) the initial sample of 12,000 youth between the ages of 14 and 21 will be drawn for the survey.

During the first year of the survey, households containing selected respondents will be reenumerated and locating information will be collected for each respondent, to aid in locating during future waves.

Upon completion of the first wave, a self-administered survey of schools will be done, to collect data about the school experience of the panel respondents.

In this presentation, I want to discuss three processes that constitute major tasks during any survey process, and how data processing might help to accomplish them; the discussion, especially the last part, is of general nature and may not be adopted as described for the NLS.

1. Instrument development
2. Controlling the survey process
3. Establishment of a data base for public use

1. Instrument Development

The questionnaire development for the NLS was done using an interactive text editor, in our case, WYLBUR. Using WYLBUR not only helped in achieving a correct instrument faster, but also and more importantly, put the questionnaire online for codebook construction and the final documentation.

A simple system was developed that uses the questionnaire as input to establish an initial codebook. This codebook contains a variable name, a variable label, a precoded values, and value labels.

The initial codebook will be used by the coding department during the editing and coding process. Each coder, as he/she works on the completed questionnaire, will be able to inspect existing codes, update, or add new codes using an interactive update system. This means that the online codebook file will contain, at any given time, all the codes that were coded in the completed questionnaires and are going to be data entered.

In the next step, the system will process the complete codebook file and create control cards for a statistical analysis system, in this case SPSS. Since coding of the questionnaires and updating of the codebook file will occur simultaneously, all codes data entered from the completed questionnaire will also appear in the codebook file, and subsequently in the control card file created from it. From the control card file and the data file frequency tables will be generated by the analysis system. These will be saved on an online device.

Going back to the original questionnaire file, it will be possible to copy the actual questions into the frequency tables, which will result in a fully documented codebook in machine readable form, available as soon as data entry is completed. At any time during the process, it will be possible to generate a partial codebook, since the codebook file will at all times be synchronized with the data coded and data entered.

2. Controlling the Survey Process

For large surveys, controlling the survey process is not a trivial task, and a system is needed that can collect data on each case throughout the entire survey and also keep track of accumulated costs.

A new survey control system was developed for NORC. It has the following objectives:

Generality. The system should be a general purpose system that can be adapted easily to any in-house study.

Ease of setup and use. Ultimately a person who is not a programmer should be able to set the system up for a new study, and data entry should be done by a clerk rather than by a programmer.

Interaction with existing NORC software. Already in existence at NORC are software systems that collect data on interviewer performance, a payroll system that collects data about the time and money spent for each survey, and a sampling system that is used to select the cases for the survey. Since the survey control system needs some of this data, it should interface with the already existing software and access the files generated by it.

Interactive updating facilities. For large surveys an interactive updating facility becomes necessary to keep up with all the activities that have to be entered into the system. Also the ability to inspect a single case in direct access mode is important to spot certain problems.

Interviewer reports. The system should provide reports that aid the data collection efforts, such as reports to interviewers to inform them about their assignments and later to keep them up to date about their outstanding cases.

Management reports. One of the major tasks of the control system is to provide management with a tool to control the progress of the survey process. It should be possible to generate at any time reports showing the completion rate, history reports showing the flow of the progress since the study began, and performance reports providing information about the performance of each interviewer and the actual cost per case.

The system that was developed, and has already lived through its first application, runs partially in interactive and partially in batch mode. Updating is done mainly in interactive mode and creation of files and reports is done in batch mode. There are functions that overlap, such as certain short reports that can be run in both modes, and almost all updating functions can be performed in the batch mode also.

The interactive part of the system uses basically two files, the so-called event file and the interviewer file.

The event file is an indexed sequential file containing one record for each case in the study, and carrying in addition to control fields, the current disposition of each case. Since the number of cases may become very large, it is essential that the record length of the event file be kept as small as possible. The event file is created by the system at the beginning of the survey from a file generated by the NORC sampling system.

The interviewer file is a direct access file containing one record for each interviewer working on the survey. It is a subset of a general NORC interviewer file, which contains all interviewers working for NORC.

The current functions of the interactive systems are:

Updating of event and interviewer file
  Case dispositions
  Control fields
  Initial interviewer assignment
  Interviewer reassignment
  Adding new cases to the event and interviewer file
  Deleting cases from the event file
Listing facilities
Status summary report

In addition to report generation, the batch part of the control system performs:

File creation
File backup
Creation of history files

The following reports have been implemented up to date:

Interviewer assignment report--informs each interviewer about the initial assignment
Outstanding cases report--gives each interviewer a list of all outstanding cases
Labels--for mailing purposes
Status summary report--a management report that tabulates the number of cases in each disposition
History report--indicates changes over time in case dispositions
Performance report--gives information about the performance of each interviewer and calculates the per case cost

The survey control system is still in its development phase. Its feasibility was tested during the survey pretest. The system met most of the design goals, but improvements have to be made to allow for greater flexibility and for a more general report-writing facility.

3. Data Base Considerations

The following discussion outlines general ideas on data base considerations and data distribution which may be applied for complex data files resulting from surveys like the NLS.

A panel study is normally viewed as administering the same instrument (questionnaire) periodically to identical respondents. In reality, however, not only does the number of respondents vary for each wave, as respondents drop out for one or more waves or entirely, but also the instrument changes over time, as some questions are dropped and others are added.

Viewing the questions as variable descriptors and each respondent as a unit of analysis (case), the data base changes in both directions.

Normally data files resulting from a panel study are distributed as rectangular files where data for each new wave is added on horizontally (by respondent) to the existing data. Data for respondents who dropped out are left blank (missing). Linkage of identical data items (same questions) from one wave to the next can only be achieved by tracing them with the use of a hardcopy codebook.

Three arguments can be made against the distribution of one general public use file, which contains all data of the data base to all users.

First, as data files tend to increase in size, the number of tape reels to be shipped to each user increases and it is hard for the user to manipulate and extract the data he is interested in.

Second, as the structure of the data bases become more complex, it might be impossible to construct a single rectangular file for public use. The alternative would be to send software to access the data base along with the data, but this is impractical because of compatibility reasons.

Third, the data base for the NLS is of such a general nature that it will attract a wide variety of users, where each might be interested in a subset of the data base only.

Today with the availability of Data Base Management Systems and Computer networks a new and more efficient method can be utilized without losing any one of the advantages of the old approach.

Currently no general purpose DBMS is available at the University of Chicago. However, NORC is planning to purchase a DBMS appropriate for data derived from social science data collection. In selecting a data base management system the following criteria seem to be important:

- Ease of data base creation
- Built-in cleaning facilities, such as range checks, consistency checks, and recode facilities
- On-line file manipulation (updating, retrieval) with data security to protect critical data from retrieval or modification
- Extraction of data from one level or across levels (upward and downward aggregation of data) and creation of a rectangular file
- Linkage of the existing data base with other files (file merge)
- Creation of new variables
- On-line updating, retrieving, and programming facilities
- Creation of data (system) files for statistical analysis systems
- Utilities (backup procedures, restoring of data pürging files, dump of data, etc.)
- Report generating facilities

Once a data base system is selected, the data base will be set up and maintained at one central site, the data distribution center. Distribution of the data might be done in various ways:

- The way it is done right now: construction of a public use file containing all the data in the data base
- Construction of special rectangular user files for distribution on tape
- Through a network the user can access the data base on a read-only basis (where certain variables might be restricted from being read). In this way the user has access to the entire data base and its documentation and can build his/her own analysis file and create new variables, using either software available at the central site or running his/her own programs.

The latter approach provides all the advantages of having the entire data base available to any given user without the burden of coping with the physical size and logical complexity of the data base.

Because data base remains in one location the user always accesses an up-to-date copy of it. Updates to the data base do not have to be distributed periodically to each individual user in the form of update tapes or completely new public use tapes. A report or notice file could be maintained to inform each user about recent updates to the data base when he/she signs on. Users are free to create their own files for data analysis, which can be stored at the central site or at the user site, or to obtain a file for data analysis that is not affected by ongoing updates to the central data base.

The idea of a central data base center is not new. In the past data archives have been set up as centers for data distribution. The new approach, however, takes advantage of the existence of data base management systems and computer networks and permits much more flexitiblity for each individual user. With a terminal in his/her office or home, the user would have all the facilities of the computer at the central site available for his/her analysis.