Preparing for data entry

By Pamela Narins, Planning Manager, J. Walter Thompson

Now that your surveys are returning, you may ask yourself, “Now what?” Well, there are two steps to complete before you can analyze data: data entry and data cleaning. This column focuses on data entry.

Shifting perspective
It is time to change your thinking. Until now, you have seen your survey as questions and answers. Now, begin to think of these elements as variables and values.

Each variable is assigned a name. Each possible response, or value, is assigned a number. You can keep open-ended or verbatim responses as text.

Assigning a unique identifier variable
The first thing to do when you receive a completed questionnaire is to assign a unique “case” number to it. Because each questionnaire is called a case, this number is often assigned the variable “caseid,” for case identification.

The importance of the caseid becomes clear the instant you question the accuracy of entered data. Without this unique number, you have no way to tie your data back to the original documents, and therefore, no way to identify or correct data entry errors.

Also, some data handling and analytical tasks are possible only if you have a way of identifying individual cases.

Choosing variable names
For this and the following sections, let’s use the example of a two-question survey. Respondents were asked their gender and height. Sometimes, when naming variables, you use numbers such as “Q1.” This can be useful for identifying a variable’s position on the questionnaire. Whenever possible, however, try to choose a name that gives you clues as to the content of the variable.

Let’s call the first variable “gender,” and the second “height,” in inches. Also, add variable caseid.

Creating values
Gender gives you two response options: male and female. For our example, assign males a value of “1” and females “2.” It is important to enter a number, rather than a letter, because many analytical procedures cannot be done with letters of the alphabet.

For a response that is already a number, simply enter that number as the value. For example, a height response of 72 inches, enter “72.”

Determine variable widths
The next step is determining a maximum character width, or amount of column space, for each variable. For example, the gender variable is one character wide, because you can only respond with one-digit numbers. The maximum width for the height variable should be two characters.

The width for caseid is dependent upon the number of questionnaires you expect to be returned. If you sent out 900 questionnaires, you should assign a width of three columns to caseid. If you sent out 2,000, the caseid should be four columns, just in case more than 999 are returned.

The variable width is important because the computer reads your data file as a string of numbers. You need to tell SPSS which numbers mean caseid, height and gender. The only way to do so is to tell the computer how long each number is.

Handling missing values
Respondents leave questions blank for many reasons. Regardless of the reason, you should key in a number for “refused” or “no answer” rather than leaving a blank. For example, if a respondent indicated they didn’t know how tall they were, you may key in a “98.”

The code “98” only works if you know no respondent is legitimately 98 inches tall. If you are not sure, then you may want to assign a value “998” to the “don’t know” response. If you do this, increase the variable width to three columns.

Whatever values you assign to missing values, be consistent. A consistent scheme makes analysis much easier to perform and interpret.

Entering data
Data entry is highly tedious, prone to error and critically important. There are several ways to get your data into the computer:

- Have SPSS interpret strings of numbers from a word processor
- Enter data using SPSS Data Entry II to create SPSS data files
- Use a spreadsheet
- Enter data directly into SPSS through the Data Editor

Other data entry methods include using a Computer Assisted Telephone Interview (CATI) system that enters data directly into a computer during the interview. You can also automate your data collection using a product like Teleform 5.0 which employs scannable forms (see Teleform article on page 1 of Keywords.)

Regardless of the method, keep in mind, errors are inevitable. If your data are carelessly entered, you can lose all the hard work that went into ensuring a sound sampling frame, usable questionnaire design and proper administration.

One safeguard you can use during data entry is called double entry. Often times entering cases and checking them twice is impractical. It is, however, reasonable to check a random sample of cases by double entry. If you have a large sample that requires multiple data entry personnel, create a variable for that person’s initials. That way, you can track the clerk’s accuracy rates.

Next time
Once you enter your data, you need to clean, or proofread, your data file. In the next issue we’ll discuss the various steps involved in data cleaning.