

Running head: CODING DATA IN MEASUREMENT AND EVALUATION

Coding Data

H. Paul LeBlanc III

Southern Illinois University at Carbondale

Manuscript submitted for inclusion as an appendix in:

Sarvela, P. D., & McDermott, R. J. (in press). Health Education Evaluation and Measurement: A Practitioner's Perspective (2nd Ed.). Madison, WI: Wm. C. Brown & Benchmark.

Abstract

An important step in conducting quantitative research involves the coding of data into a matrix of numbers which can be manipulated for analysis. The method for coding data is dependent upon:

(a) the type of analysis to be used to analyze the data, (b) the type of data, and (c) the

requirements of the statistical program to be used for the analysis. This essay describes issues to

consider when coding data for research purposes. Specifically, this essay describes the coding of

data into row by column matrices for use by computer statistical programs, and provides an

extended example from an actual health care study.

Coding Data

All research requires several systematic steps. The steps involved in conducting a study include the initial inquiry within a domain of research, a review of the literature, the formation of research questions and hypotheses, the development of the method of inquiry and testing, the operationalization of concepts, construction of the instrument, collection of data, coding of data, running of appropriate tests, analysis of results, and the writing of the report to disseminate the results. Perhaps the most tedious of these steps is the coding of the data. Careful planning in the development of the method and the construction of the instrument can eliminate many difficulties in data coding. However, the coding of data also involves procedures associated with the tools or statistical package to be used for the analysis. The purpose of this appendix is to describe the issues to consider when coding data. An extended example from an actual study using SPSS® for Windows® version 7.5.2 will be used for illustrative purposes.

Data Types, Coding and the Relationship

Between Coding and Types of Analyses

The first task after collecting data and before analysis can begin involves the coding of data into a form that can be read by a statistical program. The coding of data often requires the transformation of data points on the research instrument into a set of numbers. The method for selecting an appropriate number to represent a data point is dependent upon: (a) the type of analysis to be used to analyze the data, (b) the type of data, and (c) the requirements of the statistical program to be used for the analysis.

The type of analysis to be used to analyze the data is determined typically before the questionnaire is developed and is specified by the structure of the research questions or hypotheses. If the research questions are posed to determine a relationship between variables,

those variables must be operationalized by the instrument in such a way as to allow for correlational analysis. Likewise, if hypotheses are posed to compare the means or variances between two or more groups of respondents, the data for the two groups must be coded similarly to allow for direct comparison. For example, if the researcher wishes to compare the attitudes of male and female students toward the use of inferential statistics in public health research, then the attitudinal variable should be coded in such a way that an independent mean can be derived for each of the two groups, male or female.

The operationalization of the research questions, therefore, affects the choices of the types of analyses to be run, which in turn affect how the data should be coded. Depending on the research questions, nominal, ordinal, interval or ratio level data may be collected with the instrument. In the case of survey research, many different types of questions may be asked which can be coded as any of the four data types. Regardless of the type of data, the data must be coded in such a way as to allow for the distinction between data points by case and variable. The most efficient means of coding is to create a “case-by-variable” matrix.

Statistical Packages and Coding Needs

For most statistical programs that data matrix will require individual cases to be set in a row, and variables for each case to be set in columns. For example, for each individual subject (each returned questionnaire), a row of numbers representing their responses will be utilized. Such a case might be represented as below:

```
001 21 23 25342 55232 34244
```

In this example, the first three columns (the number 001) might represent an identification number for subject number one. The second number 21 might represent the age of the subject. The third and fourth numbers (2 and 3 respectively) may represent demographic variables such as

sex and year in school. The next three sets of numbers may represent the dependent variables from the survey instrument. For each of the three sets, each individual number may represent a response to the questionnaire. In this example, the data has been coded in sets of five numbers with a space between in order to assist the researcher in finding errors in the data more quickly. The three sets of five numbers in the example may represent fifteen questions on the questionnaire with each individual single digit representing a data point for case 001.

The example above assumes that each single digit represents the data point for the case with the exception of the identification number and the age. However, the number of digits required for any given variable is dependent upon the number of possible responses for any given question on the questionnaire. If for example a question is posed in the form of a Likert-type five-point scale, the possible responses for that variable would be one through five. Therefore, that data point could be represented by single digit. Missing data may be represented by the number zero (as in column "C" of case 002 below). Ratio level data, such as the age of the subject, will most likely require two digits or possibly three. As cases are added in subsequent rows, columns become more apparent as represented below:

	C					
001	21	23	25342	55232	34244	
002	20	22	35033	54332	34354	
003	20	12	34232	44233	13445	

The column identified with the letter "C" in this example would represent the third survey question variable following the case identification number (001) and the demographic questions (age, sex and grade).

Finally, consideration must be given to missing data and data recoding. Missing data occurs when a question is skipped or otherwise not answered on a questionnaire. The most

common practice for dealing with missing data is to use the number zero as a place marker in the data matrix. Place markers allow the row length to be consistent for all cases and also insure that variable columns maintain their integrity. In some instances, however, zero may not be the best choice as when zero has a particular meaning. Whatever number is chosen as the missing data place marker, that number should be used consistently and must be specified as representing missing data.

On occasion data may need to be recoded. Data recoding refers to the procedure of specifying numeric representation of data points which are somehow different than what is indicated by the instrument. For example, it may be necessary to phrase some questions negatively and others positively on the questionnaire in order to reduce the possibility of response bias¹. A negatively coded Likert-type scale item may have “1” equal always and “5” equal never, whereas a positively coded Likert-type scale item may have “1” equal never and “5” equal always. It may be more efficient for analysis, however, to have the data recorded all positively. The data may be coded and inputted as is on the questionnaire and the recoded with a command in the statistical program. Or, the data may be inputted into the matrix or table already recoded. Either procedure is appropriate; however, the researcher should take care to note the recoding of data in the analysis and interpretation of results.

The requirements of the statistical program will specify how the data should be coded. The above three case sample might be used in statistical programs that can read straight text such as an ASCII text file. Many statistical packages will read data in this form. The two most widely used statistical programs, SAS[®] and SPSS[®], read data in a row by column matrix. These two programs however access the data differently. In text-based systems, such as mainframe, UNIX or DOS-based computer systems, these programs require that the analyst write a batch program

which accesses the data. In SAS®, the data are included internally in the batch program. In SPSS®, the data reside in a separate file which is called from the batch program. Below is an example of a SAS® batch program for describing the means of the first three dependent variables with the data in the above example.

```

title 'Class attitudes toward statistics';
data classatt;
  input case 1-3 age 5-6 sex 8 grade 9 A 11 B 12 C 13 . . .;
  cards;
001 21 23 25342 55232 34244
002 20 22 35033 54332 34354
003 20 12 34232 44233 13445;
proc means data=classatt;
  var A B C;
run;

```

SPSS® uses similar batch program language for text based systems. Below is an example of an SPSS® batch program which describes the means of the first three dependent variables. In this example, a recode command has been used to change negatively ordered Likert-type scale item (column “C”) to a positively ordered variable. As well, a missing values command specifies the number zero.

```

TITLE 'Class attitudes toward statistics'.
DATA LIST FILE='classatt.txt'
  /ID 1-3 AGE 5-6 SEX 8 GRADE 9 A TO E 11-15 . . . .
RECODE C (1=5) (2=4) (3=3) (4=2) (5=1).
MISSING VALUES ALL (0).
FREQUENCIES VARIABLES=A TO C
  /STATISTIC=MEAN.
FINISH.
//.

```

The SPSS® batch program calls the file 'classatt.txt' which contains the data in the form:

```
001 21 23 25342 55232 34244
002 20 22 35033 54332 34354
003 20 12 34232 44233 13445
```

In these two examples, both SPSS® and SAS® require that data be coded in a row by column matrix and saved as an ASCII text file. Both SPSS® and SAS® now offer statistical programs that work in the graphical Windows® environment. Although these programs can still use data saved in a text file, they now provide the user with the ability to input data into a table or spreadsheet. Such a table separates columns and rows. This format greatly improves the ease of coding data. Below is an example of a table using the same data as in our previous example (see Figure 1).

	age	sex	grade	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	21	2	3	2	5	3	4	2	5	5	2	3	2	3	4	2	4	4
2	20	2	2	3	5	0	3	3	5	4	3	3	2	3	4	3	5	4
3	20	1	2	3	4	2	3	2	4	4	2	3	3	1	3	4	4	5

Figure 1. Data input table in a graphical environment.

One of the benefits of using a table is the ability to list the variable names at the top of each column. Furthermore, since each row is numbered, an identification column is no longer necessary. Tables also make it easier to check the accuracy of the inputted data against the questionnaire. Tables allow the user to select a single or group of columns or rows for data cleaning (error correction) or manipulation. In general, tables can be moved from one computer application to another without jeopardizing the integrity of the data. In the following extended example, tables were used for the development of a codebook⁴ as well as data input into a “case-by-variable” matrix.

An Extended Example from an Actual Study

In a study conducted by LeBlanc and colleagues⁵, we distributed a survey questionnaire to a sample of physicians and certified nurse-midwives in medically underserved areas throughout Illinois and Indiana. The purpose of the study was to determine the professional climate for collaboration between these two groups of medical professionals. A questionnaire was developed which collected nominal, ordinal, interval and ratio level data. The questions were designed to determine medical practice characteristics and attitudes of both certified-nurse midwives and physicians. Selected questions from the survey can be found in Figure 2 (see below).

Data types in the survey questionnaire. The response to questions regarding age (1), time in practice (2), and number of deliveries (5) are ratio level in nature. Questions three and four ask respondents to specify the nature and scope of their practice. These data are nominal in nature as they name different practice contexts. Questions six and seven are attitudinal Likert-type scales which are treated as interval level data. The possible responses for these two questions can be averaged for comparison between groups such as between physicians and certified nurse-midwives. Finally, question eight was developed as a ranking question in which respondents were asked to rank order responses. Because no set interval exists between ranked items, data associated with question eight were treated as ordinal level data.

Coding of the survey questions. For the nominal level data, responses had to be coded as numbers to conduct statistical analysis. Although numbers were already associated with the interval and ordinal level data, the values associated with those numbers had to be specified for analysis. Ratio level data are pure numbers that do not have to be interpreted or coded but simply are recorded as is. However, these data points must be entered into the data matrix table in order to be used in the analysis.

Survey No. P- _____
1. What is your age? _____
2. How long have you been in practice (not including residency)? _____ Year (s)
3. How would you best describe your practice?
a. Solo practice
b. Group of MD's practice
c. Group of MD's and nurse-midwives
d. Work in a hospital
e. Work exclusively for an HMO
f. Work in an academic or teaching setting
g. Work in a Federal, State or Local agency
4. What is the scope of your practice (related to maternity)?
a. Complete obstetric and gynecological services, including deliveries
b. Antepartum, postpartum, and family planning services
c. Other limited services, including education and training
d. Do not provide maternity related services
5. Approximately how many deliveries did you attend during 1995? _____
6. Do you believe certified nurse-midwives provide an acceptable alternative to physician care for women with low risk pregnancies?
_____ 1 2 3 4 5 _____
never sometimes neutral often always
7. Do you believe certified nurse-midwives have the skills necessary to evaluate between low and high pregnancy risks?
_____ 1 2 3 4 5 _____
never sometimes neutral often always
8. What are your reason(s) for consulting with or receiving referrals from this nurse-midwife/group? (please rank all that apply, using 1 to indicate most important, 2 to indicate next, etc.)
_____ The hospital where I practice requires this contact
_____ It increases referrals of patients to my practice
_____ It helps meet the needs of my community since there are too many obstetric patients to be seen by the physicians alone
_____ It helps meet the needs of my community because a certified nurse-midwife provides access to obstetric care for patients who can not afford a physician's care
_____ Just to assist when problems arise
_____ Meet needs of patients who prefer nurse-midwifery care
_____ Other _____

Figure 2. Selected survey questions from the LeBlanc, et al. (1997) study.

Question eight also provided for an open-ended response with an “other” category.

Open-ended questions cannot be coded prior to the return of all questionnaires, but rather require

the researcher to do some initial content analysis to determine the plausible categorization of data for coding. Open-ended questions allow the researcher to discover the perspectives or priorities of the respondent without imposing a limit on possible responses. However, care must be taken to insure that the respondent's intent is captured. Therefore, all responses to open-ended questions have to be listed and then reduced to common themes.

Item	Variable Name	Variable Label	Value Labels
	id	Case Identification Number	
1	age	Age of respondent	(ratio level data)
2	practice	Time in practice	(ratio level data)
3	type	Practice arrangement	1 = a, 2 = b, 3 = c, 4 = d, 5 = e, 6 = f, 7 = g, 8 = h
4	scope	Scope of practice	1 = a, 2 = b, 3 = c, 4 = d
5	delivery	Number of deliveries	(ratio level data)
6	alternat	CNMs acceptable alternative	1 = never to 5 = always
7	risk	CNMs have necessary skills	1 = never to 5 = always
8a	reason1	MD hospital requires cont.	0 = Not selected, 1 = Selected
8b	reason2	Increase referrals to MD	0 = Not selected, 1 = Selected
8c	reason3	Meet needs of comm. acces	0 = Not selected, 1 = Selected
8d	reason4	Meet needs of comm. econ.	0 = Not selected, 1 = Selected
8e	reason5	Assist when problems arise	0 = Not selected, 1 = Selected
8f	reason6	Meets need of patients	0 = Not selected, 1 = Selected
8g	reason7	Other reasons listed	1 = CNMs require MD backup, 2 = Large number of patients, 3 = CNMs give quality care

Figure 3. Data coding sheet for selected survey questions for the LeBlanc, et al (1997) study.

Development of the codebook. To assist in the inputting of data into the “case-by-variable” matrix (table), a codebook was developed which specified the variable name, variable label, and possible values associated with each item on the questionnaire. The codebook indicates how data points are to be inputted into the data matrix. Figure 3 shows the codebook developed for the data coding task of the LeBlanc, et al. (1997) study.

For example, on the questionnaire item three asks respondents to circle the appropriate letter. In the codebook, we specified by which number each possible response was being represented. For items six and seven, we specified the range of possible responses, and for item 8g

we specified what other responses were offered by respondents to the open-ended “other” question. The “ID” variable corresponds to the “Survey No. P-___” box on the top right corner of the questionnaire. The purpose for this box is to be able to check the questionnaire responses against the inputted data for accuracy.

Items 8a through 8f were listed in the codebook with values of zero or one. Although the questionnaire asked respondents to rank order the importance of their reasons for consulting with certified nurse-midwives, very few respondents actually did. The vast majority of respondents placed a mark next to those reasons with which they agreed. The coding scheme we derived for those items will allow us to determine a rank order based on the frequency a given reason was marked. In this sense, the codebook does not agree exactly with the structure of the questionnaire. In the future, revisions to the questionnaire will ask respondents to check all that apply. However, the coding scheme that was developed for this study and was placed in the codebook will assist in the next step: inputting of data.

Data coding in SPSS®

With the codebook in hand, researchers can have assistants help with the inputting of data. As discussed above, the inputting of data can be accomplished by creating a text file with rows of numbers for each case. In SPSS® for Windows®⁶, data can be inputted directly into a table. Figure 4 below shows the data editor screen in SPSS® which is used for inputting data.

In this figure the cases and variables are evident with each row representing a case and each column representing a variable. The variable names are listed across the top above each column. The variable names on the data editor screen correspond to the variable names listed in the codebook. SPSS® for Windows® also allows the researcher to define the variable characteristics including the variable label and value labels. For example, for questionnaire item

four, the variable name defined in SPSS® is “scope”. The variable label for “scope” is “Scope of Practice”. Value labels are assigned to each of the possible responses for “scope”. These labels link a number to the actual response on the questionnaire. This ability to correspond variable and value labels assists in the interpretation of results as SPSS® prints the labels in the output.

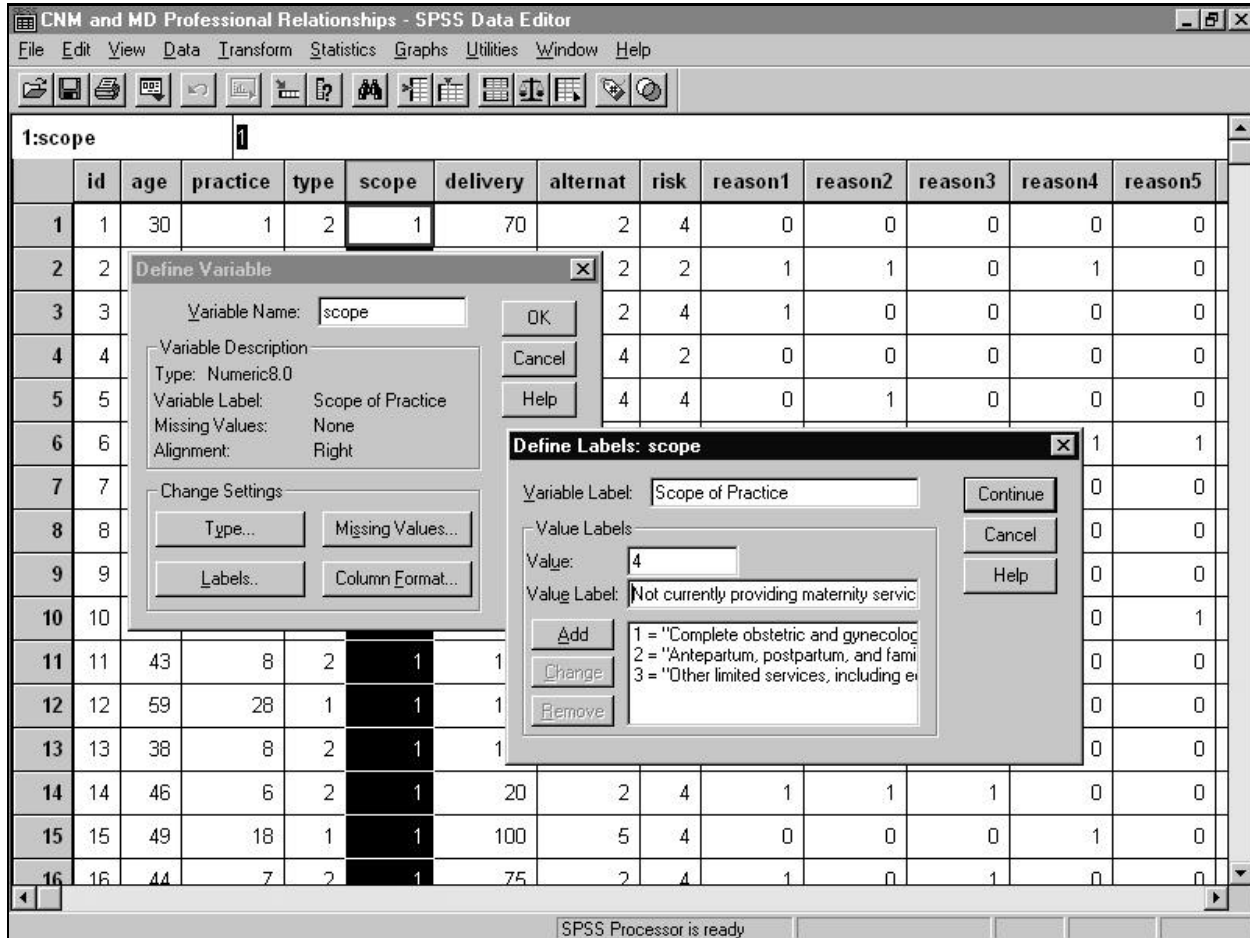


Figure 4. SPSS® for Windows® version 7.5.2 Data Editor screen for LeBlanc, et al. (1997) study.

To define the variable, a column is selected which corresponds with the placement of the variable according to the instrument. In our example, the fourth column after "ID" corresponds to the fourth question on the questionnaire. The first command under the data menu allows the researcher to define the variable characteristics. Those characteristics include the type of variable, the variable label, the missing values, and column formatting. All of the data in the sample was coded as numeric. SPSS® assumes numeric data. To define the variable and value labels, the labels

setting must be changed. The "define labels" dialogue box allows the researcher to define both variable and value labels. This procedure in SPSS® can be followed for all variables.

As discussed previously, the ID column may not be necessary due to the numbering of rows in the input table. In the example study, the ID column was necessary because separate questionnaires were sent to physicians and certified nurse-midwives. Distinct questionnaires were important for this study because the context for practice differs in several characteristics between the two groups. Although the two questionnaires had many different questions, they were designed to parallel each other. Therefore, many of the questions, such as scope of practice, are phrased with slightly different but compatible responses. Other variables, such as time in practice, were directly comparable between two groups. The goal of the research project was to compare characteristics in attitudes between the two groups. Both questionnaires, the codebook, and the data matrix had to be developed holistically.

Discussion of Coding and Analysis Issues

In SPSS®, analysis of data is accomplished by selecting variables (columns) and running appropriate tests. In our study, there were several things we wanted to know. For instance, we wanted to know if there was a difference between the average amount of time in practice for physicians and certified nurse-midwives. The dependent variable for this analysis was located in the practice column. The variables chosen for analysis are determined by the research questions or hypotheses. After selecting the appropriate variables and statistical tests, SPSS® outputs the results in a second window. These results can then be saved, copied or pasted to another application such as a word processor.

One benefit of using a Windows®, or other graphical user interface, statistical program is the ability to import or export data and results between applications. This makes creating data files or reports easier for the researcher. For example, the questionnaire could be developed with

the forms engine of a database program. This is particularly helpful when data are being collected online. Data and variable labels can be imported directly into a statistical program such as SPSS® or SAS®. Furthermore, results output such as charts and tables can be exported directly into word processors or graphics programs.

These applications are a far cry from using a slide rule or punching data cards, or even writing batch programs for use on text-based computer systems. However, the newer graphical user interfaces do not absolve the researcher from the responsibilities of checking data for accuracy. A critical step in the coding of data is the comparison of the inputted data with the responses on the survey instrument. Although the use of these new tools allow research and analysis to be performed more expeditiously, the results of the analysis will be inaccurate if data are not coded and inputted correctly.

Footnotes

- ¹ Smith, M. J. (1988). Contemporary communication research methods. Belmont, CA: Wadsworth.
- ² See: Schlotzhauer, S. D., & Littell, R. C. (1987). SAS[®] System for elementary statistical analysis. Cary, NC: SAS Institute.
- ³ See: Norušis, M. J. (1990). SPSS/PC+™ 4.0 Base Manual for the IBM PC/XT/AT and PS/2. Chicago: SPSS.
- ⁴ See: Bloom, M. (1986). The experience of research. New York: MacMillan.
- ⁵ LeBlanc, H. P., III, Simon, B., Garard, D. Nawrot, R., Roberts, J., & Stiffler, D. (1997, November). Certified nurse-midwives and physicians: Identifying biases and barriers to successful collaboration. Paper presented at the 125th annual meeting of the American Public Health Association, Indianapolis, IN.
- ⁶ SPSS. (1997). SPSS[®] Base 7.5 for Windows[®]: User's Guide. Chicago: Author.