Encyclopedia of Social Science Research Methods

Coding

Contributors: Linda B. Bourque
Editors: Michael S. Lewis-Beck & Alan Bryman & Tim Futing Liao
Book Title: Encyclopedia of Social Science Research Methods
Chapter Title: "Coding"
Pub. Date: 2004
Access Date: April 06, 2015
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780761923633
Online ISBN: 9781412950589
DOI: http://dx.doi.org/10.4135/9781412950589.n128
Print pages: 133-137

http://dx.doi.org/10.4135/9781412950589.n128
Coding is the process by which verbal data are converted into variables and categories of variables using numbers, so that the data can be entered into computers for analysis. DATA for social science research are collected using SELF-ADMINISTERED QUESTIONNAIRES and questionnaires administered through telephone or face-to-face interviews, through observation, and from records, documents, movies, pictures, or tapes. In its original “raw” form, these data comprise verbal or written language or visual images. Although many data entry programs can handle at least limited amounts of linguistic data, for purposes of most analyses, these data must eventually be converted to variables and numbers and entered into machine-readable data sets. Each variable in a data set (e.g., sex or gender) must consist of at least two categories, each with a unique code, to be a variable. It must have the possibility of varying; if it does not vary, then it is a constant. Thus, a study could include both men and women, where gender or sex is a variable, or it could include only men or only women, where gender is a CONSTANT. Numeric codes for gender can be assigned in both situations.

A single unique code can consist of a single digit or multiple digits, with the functional upper number of digits being 8 or 10. A code consisting of multiple digits may have multiple variables embedded within it. For example, respondents were asked after earthquakes whether various utilities were unavailable, including water, electricity, gas, and telephones. Four two-category variables can be created for each utility, using 1 for “yes, it was off” and 2 for “no, it was not off.” But the four single-digit variables can also be combined to create a four-digit variable that creates all the possible combinations of the four variables. An example would be the code of 1221, which would indicate that water and telephones went off but electricity and gas stayed on.

The numbers assigned can be meaningful in and of themselves, or they can function simply as a shorthand representation of the verbal data. If, for example, data on age are collected, the age recorded for each person, institution, state or country, or object is meaningful, and statements can be made about the relative ages of people, states, or objects. If, in contrast, data on religious affiliations of people, households, or institutions are collected, the number assigned to each religion is arbitrary—for example, 1 for Catholics, 2 for Buddhists, and 3 for Hindus.

Levels of Measurement

In developing codes for verbal data, the concept of LEVELS OF MEASUREMENT is often used (Stevens, 1968). There are four levels of measurement: NOMINAL, ORDINAL, INTERVAL, and RATIO. As we move from nominal data to ratio data, the computations and statistics that can be used during analysis become more diverse and sophisticated. When creating codes for data, it is important to understand and consider these differences.

Codes for religion are considered nominal. Nominal codes are arbitrary, shorthand space holders. In creating codes for nominal data, the researcher is concerned with creating exhaustive and mutually exclusive codes. *Exhaustive* means that a unique code number has been created for each category of the variable that may occur in the data. Thus, in creating codes or code frames for a variable on religious affiliation, it may be necessary to create unique code numbers for individuals, institutions, or groups that have no religious affiliation, as well as for those that identify as agnostics or atheists. *Mutually exclusive* means that the information being coded about each person, institution, object, or grouping can be clearly assigned to only one category of the variable. So, for example, it is clear that Buddhists always get a code of 2 and never get a code of 1 or 3.

The categories of a nominal variable have no meaningful rank order, and the distance between the categories cannot be calculated. The only measure of central tendency that can be calculated appropriately for a nominal variable is the mode. Many of the most important variables included in social science research are naturally nominal. These include sex or gender, religion, language, ethnic identification, and marital status.

Ordinal variables are considered a higher level of measurement. The categories of an ordinal variable can be meaningfully rank ordered. So, for example, a person who drinks one pint of water drinks less water than a person who drinks two quarts who, in turn, **[p. 133 ↓]** drinks less water than a person who drinks two gallons, and they can be assigned codes, respectively, of 1, 2, and 3. The person with a code of 3 clearly drinks more water than those with codes 1 or 2, but nothing can be said about the relative distance between the three people when those three codes are used in analysis.

Both modes and medians can be calculated for ordinal variables, but it is technically incorrect to add or subtract scores or to calculate mean scores. One of the problems with assigning codes to variables that are naturally ordinal is developing codes that are both mutually exclusive and exhaustive. In the above example, the codes are mutually exclusive but are not exhaustive. Note there is no code for a person who drinks no water, three gallons of water, or one quart of water. Age can be regarded as an instance of an interval variable, which requires a scale with equidistant (or constant) numerical categories. Consider the answer to the question, “How old were you on your last birthday?” With interval variables, the categories can be rank ordered, and one can talk about the equal distance between the categories. For instance, interval variables have an arbitrary zero, the categories can be rank ordered, and we can talk about the distance between categories. For example, the distance between 24 and 26 is the same as the distance between 33 and 35. In addition to the mode and median, statistics such as the mean, STANDARD DEVIATION, and CORRELATION are meaningful and can be interpreted.

Ratio variables differ from interval variables in that they have a meaningful zero and are completely continuous, meaning that every point on the scale exists. Weight and height are often considered ratio variables. Ratio variables are also called continuous variables, whereas the other three types of variables are sometimes called DISCRETE variables. For purposes of most social science research, the distinction between interval and ratio is largely arbitrary because most of the statistical analyses that can be done on ratio variables can also be done on interval variables. In contrast, analysis techniques that are appropriate for interval and ratio variables technically should not be used on nominal or ordinal variables.

Determining Codes for Variables

Codes for data can be developed during the design of the study or after the data are collected. When structured questionnaires or abstraction forms are used to collect data on a finite number of variables and most variables have a finite number of possible answer categories, codes should be developed as the data collection instrument is being finalized. When OBSERVATIONS, FOCUS GROUPS, SEMI-STRUCTURED INTERVIEWS, or DOCUMENTS are the source of data, codes generally will be

developed inductively using CONTENT ANALYSIS or other procedures after data collection is completed. The development of codes for CLOSEDENDED QUESTIONS in a questionnaire will be used as an example for the first kind of code development, whereas the development of codes for openended questions will be used as an example for the second kind.

A number of things should be considered when developing codes. As noted above, codes and the associated answer categories should be both *exhaustive* and *mutually exclusive;* that is, there should be a code for each answer that might be given, and respondents, data collectors, and data entry personnel should have no difficulty determining which answer fits the situation being described and which number or code is associated with that answer. For example, after the Loma Prieta earthquake, the authors (Bourque & Russell, 1994; Bourque, Shoaf, & Nguyen, 1997) asked respondents, “Did you turn on or find a TV or radio to get more information about the earthquake?” and provided the responses of “Yes, Regular TV,” “Yes, Battery TV,” “Yes, Regular Radio,” “Yes, Battery Radio,” and “No,” with preassigned codes of 1, 2, 3, 4, and 5. Interviewers asked respondents to select a single answer.

Although the list of answers and codes appeared both exhaustive and mutually exclusive, it was not. First, many respondents sought information from more than one form of electronic media and, second, many respondents sought information from their car radio. The question-and-answer categories were changed to add car radios to the list, and respondents were allowed to select more than one answer. If that had not been done, data would have been lost.

When multiple information is recorded or coded for what is technically one variable, each response category becomes a variable with two codes or values, mentioned or not mentioned. In the example, what started out as one variable with five possible answers became six interrelated variables, each with two possible answers or codes.

Residual other categories provide another way to ensure that the code frame is exhaustive. “Residual other” is used to deal with information that the researcher did not anticipate when the code frame was **[p. 134 ↓]** created. In the same study, the author provided a list of ways in which houses might have been damaged by the earthquake, including collapsed walls, damaged roofs, and so forth. What was not anticipated were

the elevated water towers that are used by residents of the hills outside Santa Cruz, California. Because a residual other was included, when respondents said a water tower had been damaged, interviewers circled the code of 19 for “Other” and wrote “water tower” in the space provided, as demonstrated here:

Other..... 19
SPECIFY: _____

When data are entered into the computer, both the code 19, which indicates that an unanticipated answer was given, and the verbal description of that answer, water tower, are typed into the data set.

Reviews of past research and PRETESTING of data collection procedures can help researchers determine if the answer options and associated codes are both mutually exclusive and exhaustive, but sometimes social change will cause well-tested code schemes to become inaccurate, incomplete, or even completely obsolete. For example, the standard categories used to code marital status are “never married,” “married,” “divorced,” “separated,” and “widowed,” but increasingly, both different- and same-sex couples are cohabiting. According to the 2000 U.S. census, 5.2% of households in the United States currently describe themselves as “unmarried partner households.” In a study of social networks, for example, failure to consider and account for this change may result in serious distortion of data.

Deciding on the categories and codes to use for income, education, and age is another place where problems occur. Generally, income is coded as an ordinal variable. Categories are developed such as the following: less than $10,000, $10,000–$19,999, $20,000–$29,999, and so forth. The problem comes in deciding how broad or narrow each category should be and what the highest and lowest categories should be. If categories are too broad, the top category set too low, or the bottom category set too high, heaping can occur. *Heaping* is when a substantial proportion of the data being coded falls into a single category. For example, if undergraduate college students are being studied and age is set in ordinal categories of 18–21, 22–24, and 25–29 with associated codes of 1, 2, and 3, the overwhelming majority of students will be 18 to 21, and they will be assigned a code of 1. When this happens, the variable becomes useless because it no longer has any VARIANCE.

Another problem with creating ordinal variables out of data (e.g., age, education, and income) that are naturally at least interval in underlying structure is that it may restrict data analysis. If age is coded into categories of 18–24, 25–29, 30–34, and 35–39, and so forth, the researcher can determine what the median age category is but can never calculate the mean age. Furthermore, writing out the age categories in a questionnaire or data collection form takes substantially more room than simply asking for age at last birthday or birth date. In this case, a *short open-ended* question in a data collection form may obtain just as much data—data that are just as valid and, at the same time, yield data that are more easily analyzed.

Similar problems occur when researchers use answer categories such as “none,” “some,” and “a lot.” Although *none* is pretty clear, what do *some* and *a lot* mean, and how should they be coded? If this is a question about how many people a respondent knows who were injured in a flood, why not ask instead, “How many people in all do you know who were injured in the flood?” The information collected will result in data that more closely resemble interval data, and it will not be subject to the variety of ways in which terms such as *some* and *a lot* will be interpreted.

When coding data, it is as important to record that information is missing as it is to record the data that are there. Standard MISSING DATA conditions include “not applicable,” “don't know,” “refused,” and “missing.” Codes for “don't know” and “refused” are more frequently used when data are collected directly from people, whereas codes for “missing” are extremely important when data are being abstracted out of records or other documents. “Not applicable” codes are used in both situations. “Not applicable” is appropriately used when some data are not relevant in certain situations. Often, the need for “not applicable” occurs because the data being coded are dependent on some prior information that results in the current data being irrelevant. For example, men are not asked about how they felt when they were pregnant, and it is irrelevant to look for hemoglobin scores in medical records if no blood was drawn.

Often, when data are being abstracted out of documents or records, data instructors are instructed to look for particular kinds of information and then record it using a code. If CONTENT ANALYSIS is being used to code the content of the front page of eight major **[p. 135 ↓]** U.S. newspapers over a period of a month, and data collectors are supposed to record all articles that reference the president and code for the “tone” of

the article (e.g., whether it is complimentary, critical, analytical, or whatever), the fact that no reference to the president occurs on a given front page is important information. Hence, a code for missing is included in the code frame or range of numeric codes being used to represent the content of the page. The code for missing provides at least indirect evidence that the data collector looked for information to code and did not find it.

The timing of and extent to which codes for “don't know” should be used are one of the more controversial areas of data collection. The meaning and value of including a “don't know” category and code differs with the following factors: the researcher's objectives, how data are collected, whether factual or attitudinal information is being solicited, whether respondents are told that a “don't know” category is available, and the extent to which data collectors are trained and monitored. When data are collected directly from people and the primary objective is to collect factual or behavioral information, it is better not to include a “don't know” in the available answer categories during data collection because interviewers and respondents will be tempted to select “don't know” when other answers are more appropriate. Complicating this is the fact that in interactions such as interviews, people often use phrases such as “don't know” while they think about a question. In such cases, “don't know” is an acknowledgment that the person heard the question, but it is not necessarily intended to be the answer to the question. When attitudes or opinions are being solicited, an identified “don't know” or “no opinion” answer category may be appropriate, particularly when respondents are asked their opinion about people, policies, or practices that many know nothing about.

Some categories of answers occur repeatedly throughout data collection, such as yes, no, refused, don't know, missing information, and inapplicable. Researchers can simplify data analysis by using *consistent numeric codes* throughout the data collection process for these common response options. Table 1 shows the traditional codes generally used.

Finally, researchers should consider developing code frames that *minimize the need for transformations* during data analysis. For example, if data about immunizations are being collected off medical records that represent well-baby visits, it makes much more sense to record the actual number of immunizations given at each visit rather than creating categories for “no immunizations,” “1–2 immunizations,” “3–4 immunizations,” and “5 or more immunizations” with codes, respectively, of 1, 2, 3, and 4.

		Example Code	
Response	Single Digit	Double Digit	3 + Digits
Yes	1		
No	2		
Refused to answer	7	97	997
Don't know	8	98	998
Missing information	9	99	999
Not applicable	0	00	000

Postcoding

Code frames for some kinds of data simply cannot be set up as the data collection procedures are being finalized. Open-ended questions that have no predetermined list of answer categories are included in interviews for a reason—namely, because the researcher is unable to anticipate the kinds and amounts of data that may be elicited or cannot create exhaustive, mutually exclusive codes that can easily be explained to and used by data collectors and data entry personnel. This is particularly true of exploratory research in which the objective is to find out whether more structured research is possible and would be of value. In such studies, interviews or other methods of data collection may be unstructured or only partially structured. Often, material is audio or visually recorded and transcribed prior to coding and analysis. When data collection is complete, the task is to determine what, if any, patterns exist in the data. This is done inductively through procedures such as content analysis. Similar procedures must be used to analyze documents, movies, and television programs. But the end objective is to come up with code frames or numeric representations of the data such that they can be analyzed and summarized.

Content analysis and related procedures are complex, and detailed information about how to do them is beyond the scope of this entry. Rather, some general guidelines can be provided here. First, the researcher must specify the objectives for which the code frame is to be used. If children have been observed at play, is **[p. 136 ↓]** the

objective to count the frequency of behaviors, the duration of behaviors, the intensity of the interactions, or some combination? By identifying how the data are to be used, the researcher identifies some variables that will be extracted and others that will be ignored. The process of identifying the objectives may involve preliminary gross content analysis so that overall themes can be inductively developed from the data.

Second, in developing the code frame, a balance must be realized between too much detail and not enough detail. If clothing is being described, is it important to code the number and type of buttons on the clothing, or is it sufficient to say it is a jacket? Third, and related, is maximizing the maintenance of information within the code frame as it is created. Of particular concern here is information that appears in only a few of the documents being analyzed but that is particularly pertinent to the research questions being examined. Because content analysis is a lengthy, laborintensive process, researchers sometimes fail to code information that is important because they are tired or bored. As long as the raw data exist, the researcher can always reexamine the data, but this of ten means reading many transcripts or listening to many audiotapes. Thus, every effort should be made to incorporate rare but important data into the code frame as they are being developed. This is why actively searching for information and systematically coding the fact that certain information is missing is particularly important.

Finally, researchers need to create a sufficient range of codes, variables, or dimensions so that the coder does not force data into categories that are really inappropriate. Take, for example, answers to the following statement: “Describe the vehicle(s) involved in the motor vehicle crash.” First, we could code for the number of vehicles that were involved in the crash. Then, we could code for the type of vehicle(s)—sedan, sport utility vehicle (SUV), or pickup truck—that was involved. If examining safety features is an objective of the study, it probably is important to code for the make of the vehicle, the year, and whether the vehicle had air bags. The point here is that a single stimulus—whether it is a question, a paragraph in a document, a period of observation, or a sequence in a movie or tape—may yield multiple pieces of information that become multiple variables during analysis.

Linda B.Bourque

http://dx.doi.org/10.4135/9781412950589.n128

See also

- Coding Qualitative Data

References

Bourque, L. B., & Clark, V. A. (1992). Processing data: The survey example (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–085) . Newbury Park, CA: Sage.

Bourque, L. B., & Russell, L. A. (with Krauss, G. L., Riopelle, D., Goltz, J. D., Greene, M., McAffee, S., & Nathe, S.) (1994, July). Experiences during and responses to the Loma Prieta earthquake . Oakland: Governor's Office of Emergency Services, State of California.

Bourque, L. B., Shoaf, K. I., Nguyen, L. H. Survey research . International Journal of Mass Emergencies and Disasters 15 71–101 (1997).

Stevens, S. S. Measurement, statistics, and the schemapiric view . Science 161 849–861 (1968).