a0005 Coding Variables

Lee Epstein
Washington University, St. Louis, Missouri, USA

Andrew Martin
Washington University, St. Louis, Missouri, USA

Glossary

g0030 **codebook** A guide to the database that the researcher is creating—a guide sufficiently rich that it not only enables the researcher to code his or her data reliably but also allows others to replicate, reproduce, update, or build on the variables housed in the database, as well as any analyses generated from it.

g0005 **observable implications (or expectations or hypotheses)** What we expect to detect in the real world if our theory is right.

g0010 **reliability** The extent to which it is possible to replicate a measurement, reproducing the same value (regardless of whether it is the right one) on the same standard for the same subject at the same time.

g0015 **theory** A reasoned and precise speculation about the answer to a research question.

g0020 **variable** Observable attributes or properties of the world that take on different values (i.e., they vary).

g0025 **variable, values of** Categories of a variable (e.g., male and female are values of the variable gender).

p0005 Coding variables is the process of translating attributes or properties of the world (i.e., variables) into a form that researchers can systematically analyze. The process entails devising a precise schema to account for the values that each variable of interest can take and then methodically and physically assigning each unit under study a value for every given variable.

s0005 Introduction

p0010 Social scientists engaged in empirical research—that is, research seeking to make claims or inferences based on observations of the real world—undertake an enormous range of activities. Some investigators collect information from primary sources; others rely primarily on secondary archival data. Many do little more than categorize the information they collect; but many more deploy complex technologies to analyze their data.

Seen in this way, it might appear that, beyond following p0015 some basic rules of inference and guidelines for the conduct of their research, scholars producing empirical work have little in common. Their data come from a multitude of sources; their tools for making use of the data are equally varied. But there exists at least one task in empirical scholarship that is universal, that virtually all scholars and their students perform every time they undertake a new project: coding variables, or the process of translating properties or attributes of the world (i.e., variables) into a form that researchers can systematically analyze after they have chosen the appropriate measures to tap the underlying variable of interest. Regardless of whether the data are qualitative or quantitative, regardless of the form the analyses take, virtually all researchers seeking to make claims or inferences based on observations of the real world engage in the process of coding data. That is, after measurement has taken place, they (1) develop a precise schema to account for the values on which each variable of interest can take and then (2) methodically and physically assign each unit under study a value for every given variable.

And yet, despite the universality of the task (not to p0020 mention the fundamental role it plays in research), it typically receives only the briefest mention in most volumes on designing research or analyzing data. Why this is the case is a question on which we can only speculate, but an obvious response centers on the seemingly idiosyncratic nature of the undertaking. For some projects, researchers may be best off coding inductively, that is, collecting their data, drawing a representative sample, examining the data in the sample, and then developing their coding scheme; for others, investigators proceed in a deductive manner, that is, they develop their schemes first and then collect/code their data; and for still a third set, a combination of inductive and deductive coding may be most appropriate. (Some writers associate inductive coding with research that primarily relies on qualitative [nonnumerical] data/research and deductive coding with quantitative [numerical] research. Given the [typically] dynamic nature of the processes of collecting data and coding, however, these associations do not always or perhaps even usually hold. Indeed, it is probably the case that most researchers, regardless of whether their data are qualitative or quantitative, invoke some combination of deductive and inductive coding.) The relative case (or difficulty) of the coding task also can vary, depending on the types of data with which the researcher is working, the level of detail for which the coding scheme calls, and the amount of pretesting the analyst has conducted, to name just three.

p0025 Nonetheless, we believe it is possible to develop some generalizations about the process of coding variables, as well as guidelines for so doing. This much we attempt to accomplish here. Our discussion is divided into two sections, corresponding to the two key phases of the coding process: (1) developing a precise schema to account for the values of the variables and (2) methodically assigning each unit under study a value for every given variable. Readers should be aware, however, that although we made as much use as we could of existing literatures, discussions of coding variables are sufficiently few and far between (and where they do exist, rather scanty) that many of the generalizations we make and the guidelines we offer come largely from our own experience. Accordingly, sins of commission and omission probably loom large in our discussion (with the latter particularly likely in light of space limitations).

s0010 Developing Coding Schemes

p0030 Regardless of the type of data they collect, the variables they intend to code, or even whether they plan to code inductively or deductively, *at some point* empirical researchers require a coding schema, that is, a detailing of each variable of interest, along with the values of each variable—for example the variable RELIGION of a survey with, say, "Protestant," "Catholic," "Jewish," "none," and "other" as the values. With this sort of information in hand, investigators can prepare codebooks—or guides they employ to code their data and that others can use to replicate, reproduce, update, or build on the variables the resulting database contains and any analyses generated from it.

In the section that follows, we have much more to say about codebooks. For now let us home in on this first phase—developing coding schemes—and begin by reinforcing a point suggested by our emphasis on the phrase "at some point"; namely that, in terms of research design, many steps typically proceed the step developing a coding schema, such as devising research questions, theorizing about possible answers, generating observable implications, and so on. Even when it comes to coding variables, researchers may not begin with developing a coding schema. But—and this is our chief point—they almost always perform this task during the course of a project's life. This holds for those who create databases, as well as for those who work with databases developed by others, such as the General Social Survey and the American National Election Study; that is, users need to devise a plan of their own if they desire to transform variables contained in existing databases. p0035

We also ought acknowledge at the onset that the nature of the coding task (especially its relative difficulty) varies depending on the types of variables under investigation. If we are conducting a survey of students, all of whom are between the ages of 18 and 21, then it is relatively trivial to develop a coding scheme for the variable AGE: it would take on the values "18," "19," "20," and "21." Devising the values for many other variables is not as straightforward a task. To see this, return to the deceptively simple example of the variable RELIGION, for which we listed five possible values: "Protestant," "Catholic," "Jewish," "none," and "other." This may work well for some studies, but we can imagine others for which it would not. Consider, for example, an investigation into the attitudes of individuals who belong to a Jewish synagogue wherein the researcher desires to include the variable RELIGION. Assuming that nearly all the subjects are Jews, the five values we have listed would make little sense (nearly 100% would fall into the "Jewish" category). Accordingly, the investigator would need to alter the schema, perhaps by incorporating finer divisions of "Jewish": "Jewish-Orthodox," "Jewish-Conservative," "Jewish-Reform"—or whatever schema most appropriately enables the researcher to capture the information necessary to the research task. p0040

Other problems may exist with our original values of RELIGION; for example, what of respondents who are interdenominational in their religious preferences? They will be forced to choose among the values of RELIGION or perhaps respond with "other," even if their preference is one that combines Catholic and Protestant tenets. And, for that matter, what should we make of the "other" category? Depending on the subjects under analysis, it may be appropriate (meaning that it would be p0045 an option selected by relatively few respondents) or not. But our more general point should not be missed: Accounting for the values of the variables of interest, even for seemingly straightforward ones, may be a nontrivial task.

p0050 To facilitate the efforts of researchers to perform it, we offer the three recommendations that follow: (1) ensure that the values of the variables are exhaustive; (2) create more, rather than fewer, values; and (3) establish that the values of the variables are mutually exclusive. These guidelines reflect standard practice, along with our own experience. But there is one we presuppose: Researchers must have a strong sense of their project, particularly the piece of the world they are studying and how that piece generated the data they will be coding, as well as the observable implications of their theory that they will be assessing. Because this point is obvious, if only from the brief examples we have provided, we do not belabor it. We only wish to note that an adherence to all the suggestions that follow will be difficult, if not impossible, if researchers lack a deep understanding of the objects of their study and an underlying theory about whatever feature(s) of behavior they wish to account for.

s0015 Ensure that the Values of the Variables Are Exhaustive

p0055 Our first recommendation, simply put, is that the values for each variable must exhaust all the possibilities. To see why, consider a simple example, from a 2000 study by Frankfort-Nachmias and Nachmias, in which the values of the variable MARTIAL STATUS are enumerated as "married," "single," "divorced," and "widowed." This is not an exhaustive list because it fails to include "living together but not married" and, as such, may ultimately be a source of confusion for respondents who are asked to describe their marital status or for coders who must make a decision about another's status.

p0060 To skirt the problem, investigators typically include the value "other." As a general matter, this is not only acceptable but typically necessary. It may be the case that researchers, however well they knows their project and however well developed their theory, cannot anticipate every value of a particular variable. Moreover, even if they did, it may be impractical or inefficient to include each and every one. At the same time, however, because we can learn very little from a database replete with variables mostly coded as "other," researchers should avoid the overuse of this value when it comes time to code their data. Only by having a thorough understanding of their project—or at least an understanding (perhaps developed through pretesting) sufficient enough to be able to write down an exhaustive list of the likely-to-occur values—will they steer clear of this pitfall. Pretesting or conducting a pilot study when possible and is in general a useful way to detect potential problems of all sorts—including the overuse of "other." When pretesting reveals that "the 'other' response accounts for 10 percent of more of the total responses," Shi (1997) suggested in a 1997 study (as a rule of thumb) that researchers should add new values.

Create More, Rather than Fewer, Values s0020

To some extent our second recommendation counsels "when in doubt, include a value rather than exclude it." It reinforces from our first recommendation because this will help to avoid the problem of too many "other" codes. But it suggests something else; that analysts create values that are more, rather than less, detailed. p0065

To see why, consider the example of researchers who want to explain why lower federal appellate courts (the U.S. Courts of Appeals) sometimes reverse the decisions reached by trial courts and sometimes affirm them. For such a project investigators need to enumerate the values of the variable DISPOSITION, which they could merely list as "affirm" and "reverse" because these are the dispositions that concern her. The problem here is that appellate courts do not always simply affirm or reverse; these courts have many options available to them, as Table I indicates. p0070

Even though our analysts are interested solely in a court's decision to affirm or reverse, the guideline of "creating more, rather than fewer, values" suggests that they start with all possible values of DISPOSITION (as listed in Table I). To be sure, the researchers should know which values of the DISPOSITION ought count as a "reverse" and which should count as an "affirm"; and we would require them to specify that (e.g., values 2, 3, 4, p0075

Table I Possible Dispositions in Cases Decided by the U.S. Courts of Appeals t0005

Value	*Value label*
0	Stay, petition, or motion granted
1	Affirmed; or affirmed and petition denied
2	Reversed (including reversed & vacated)
3	Reversed and remanded (or just remanded)
4	Vacated and remanded (also set aside & remanded; modified and remanded)
5	Affirmed in part and reversed in part (or modified or affirmed and modified)
6	Affirmed in part, reversed in part, and remanded; affirmed in part, vacated in part, and remanded
7	Vacated
8	Petition denied or appeal dismissed
9	Certification to another court

Source: U.S. Court of Appeals Data Base, available at: http://www.polisci.msu.edu/pljp/databases.html

6, 7 listed in Table I might be considered "reverse"). But beginning with the more detailed values has two clear advantages (both of which have even more bearing on the second phase of the coding process, discussed later). First, whoever eventually codes the data will make fewer errors. Think about it this way: If our investigators tell their coder in advance to report values 2, 3, 4, 6, and 7 as "reversals," the coder must take two steps: (1) identify the disposition by examining the court's decision and, then, (2) identify whether it is a reversal or affirmance. But if the researcher simply has the coder identify the disposition, then the coder has only one step to take. Because every step has the possibility of introducing error, researchers should seek to reduce them.

p0080 A second set of advantages accrue when the investigators turn to analyzing their data. Because they have now coded the variable DISPOSITION quite finely, they can always collapse values (e.g., they can create "reverse" from values 2, 3, 4, 6, 7 in Table I) to generate a new variable, say, DISPOSITION2, which would house the two categories of primary interest to her ("reverse" and "affirm"). At the same time, and again because they have coded DISPOSITION finely, they will be able to ascertain whether any particular coding decision affects their conclusions. Suppose, for example, that, in collapsing values of DISPOSITION, they count value 6 (in Table I) as a "reverse," even though the court affirmed in part. Because this represents a judgment on their part (although one they should record, thereby enabling others to replicate their variable) and because the converse coding (counting 6 as an "affirm") is plausible, they will be able to examine the effect of their judgment on the results. Of course, none of these advantages ensue if they initially list only two values of disposition ("reverse" and "affirm"); while researchers can always collapse values, they cannot disaggregate those coded more coarsely. This point cannot be understated; we can never go from fewer categories to many without returning to the original data source (which oftentimes, for example, in survey research, is impossible). Coding data as finely as possible allows the researcher to encode a greater amount of information.

p0085 Despite all the advantages of creating more (rather than fewer) values, limits do exist. Consider researchers who must devise values for the variable INCOME (representing survey respondents' "total family income, from all sources, fall last year before taxes"). Following the recommendation of creating detailed values might lead the investigators to ask respondents simply to report their precise income. Such, in turn, would provide them with an exact dollar figure—or the finest possible level of detail on the variable INCOME. But very few (reputable) surveyors operate in this fashion. This is because they realize that individuals may not know that exact dollar amount or may not want others to know it. Hence, rather than running the risk of reliability problems down the road, researchers typically create values that represent income categories (e.g., "under $1000," "$1000–2999," and "$3000–3999").

We can imagine other projects/variables for which our recommendation of developing detailed values would not be advisable. But, in the main and depending on the project/variable, it is a general principle worth considering. p0090

Establish That the Values of the Variables are Mutually Exclusive s0025

Under our third guideline, researchers must be sure that they have created values such that whatever unit is under analysis falls into one and only one value. It is easy to see how the failure to follow this recommendation could lead to confusion on the part of respondents and coders alike but, unfortunately, it also easy to violate it. Consider the simple example from the 2000 study by Frankfort-Nachmias and Nachmias of a variable LIVING ARRANGEMENTS OF STUDENTS for which investigators have enumerated four values: "live in dormitory," "live with parents," "live off campus," and "live with spouse." To the extent that the values seem exhaustive and detailed, this schema meets the two other recommendations but not the third—a student could "live with parents" and "live off campus," or, for that matter, live off campus with a spouse, live with parents and a spouse, or (at some universities) live in a dorm with a spouse. The values are not mutually exclusive. Guarding against the problem requires that researchers, once again, understand their project; pretesting also may be a useful step. p0095

Assigning Each Unit Under Study a Value s0030

After (or perhaps concurrently with) developing a coding schema, analysts must methodically and physically assign each unit under study a value for every variable. Doing so typically requires them to (1) create a codebook to house the schema and other relevant information and (2) determine how they will ultimately enter their data into a statistical software package so that they can analyze them. p0100

Codebooks s0035

In line with our earlier definition, codebooks provide a guide to the database that the researchers are creating—a guide sufficiently rich that it not only enables the researchers to code their data reliably but also allows others to replicate, reproduce, update, and build on the variables housed in the database as well as any analyses p0105 generated from it. Indeed, the overriding goal of a codebook is to minimize human judgment—to leave as little as possible to interpretation.

p0110 Accordingly, while codebooks contain coding schemes (that is, the variables and the values that each can take), most contain much more—including the details about various features of the research process (e.g., information about the sample and sampling procedures, data sources, and the time period of the study). We do not provide a listing of each of these components here. (Researches should investigate the Inter-university Consortium for Political and Social Research, ICPSR, and the Data Documentation Initiative, DDI, which is "an effort to establish an international criterion and methodology for the content, presentation, transport, and preservation of 'metadata' [codebooks or, in DDI's words, "data about data"] about datasets in the social and behavioral sciences." We instead focus on those relevant to coding variables: variables, values, missing values, and coding notes.

s0040 *Variable Names*

p0115 When researchers enumerate the values of their variables, those variables have (or should have) precise meanings. To return to our previous examples, investigators, may name their variable INCOME but understand that to mean "total family income, from all sources, fall last year before taxes." Likewise to researchers coding the variable DISPOSITION, this variable may signify "the ruling of a U.S. Court of Appeals."

p0120 Codebooks contain both a short variable name (e.g., INCOME) as well as the investigators' precise definition ("total family income, from all sources, fall last year before taxes"). The reason for the former is that many statistical software packages (into which researchers ultimately enter their data) still limit the length of the variable name, to, say, eight characters. Hence, to ensure that the variable name in the codebook corresponds to the variable name in the database, codebooks typically contain the abbreviated name. (Worth noting is that limits on length are becoming less of a concern in current versions of software packages, although being able to refer to variables using some shorthand is typically valuable for other reasons.)

p0125 Conventions for naming variables abound. But because other sources provide detailed discussions of this matter we need not delve into them here. What is worthy of emphasis is this advice, from ICPSR: "It is important to remember that the variable name is the referent that analysts will use most often when working with the data. At a minimum, it should not convey incorrect information, and ideally it should be unambiguous in terms of content."

p0130 In addition to the shortened variable name, researchers supply a longer, descriptive name for each variable (for INCOME, "total family income, from all sources, fall last year before taxes"), along with the form each variable takes (for INCOME, dollars). (For variables created from survey questions, the descriptive name typically is the exact question asked of respondents.) The name should convey a strong sense of the contents of the variable, and it ought be listed in the codebook and in the database (most statistical packages allow the user to enter a longer variable identifer or variable label).

Values s0045

When researchers develop their coding schema, they create values—usually in the form of labels—for each variable, such as the nine descriptive values in Table I for the variable DISPOSITION. After (or concurrently with) doing this, they typically assign a unique number to each value (as in Table I). The codebook contains both the value numbers and the value labels (e.g., in Table I, 0 = stay, petition, or motion granted; 1 = affirmed, or affirmed and petition denied; and so on); the ultimate database houses both, but it is typically the number that the coder enters for each unit of analysis. p0135

As is the case for variable names, conventions for assigning numbers to the values of variables abound. For example, values ought be convenient, intuitive, and consistent with the level of measurement. So, for example, even though the values of a discrete variable, say GENDER, could be 1010 = male and 5020 = female (because matters of size and order are irrelevant), this convention counsels that the researcher ought begin with 0 and increase by 1 (e.g., male = 0; female = 1). Starting with the lowest values with the fewest digits, however, need not mean that the researcher must sacrifice sophistication. Manheim and Rich in their 1995 article detailing how researchers interested in a finely coded RELIGION variable can devise a numbering system that is simultaneously simple, intuitive, and refined make this point nicely. Rather than merely assigning values 1–4 to various Protestant denominations (say, 1 = Baptist, 2 = Methodist, 3 = Presbyterian, and 4 = Lutheran), values 5–7 to various forms of Judiasim (e.g., 5 = Orthodox, 6 = Conservative, and 7 = Reform), and value 8 to Catholic, researchers can classify numbers in a logical, meaningful way (10 = Baptist, 11 = Methodist, 12 = Presbyterian, 13 = Lutheran; 20 = Jewish Orthodox, 21 = Jewish Conservative, 22 = Jewish Reform; 30 = Catholic). Under this system, each major category of religion (Protestant, Jewish, Catholic) receives the same first digit, with the second digit representing a subdivision. As Inter-university Consortium for Political and Social Research explains, "This type of coding scheme permits analysis of the data in terms of broad groupings as well as individual responses or categories." p0140

Ease of use, intuition, and logic also should guide the assignment of values to continuous variables. This p0145 typically means that researchers should record the original value, reserving transformations for later. For example, even if the logarithm of AGE will ultimately serve as an independent variable in the analysis, the researcher ought code the raw values of AGE and do so sensibly (if a person is 27, then the value of the variable AGE for that person is 27).

p0150 Two other rules of thumb are worthy of note. One is that wherever and whenever possible, researchers should use standard values. If the ZIP CODE of respondents is a variable in the study, it makes little sense to list the codes and then assign numerical values to them (11791 = 1, 11792 = 2, 11893 = 3, and so on) when the government already has done that; in other words, in this case the researcher should use the actual zip codes as the values. The same holds for other less obvious variables, such as INDUSTRY, to which the researcher can assign the values (e.g., 11 = Agriculture, 22 = Utilities, and so on) used by the U.S. Census Bureau and other agencies.

p0155 The remaining rule is simple enough and follows from virtually all we have written thus far—avoid combining values. Researchers who creates a variable GENDER/RELIGION and codes a male (value = 0) Baptist (value = 10) as value = 010 are asking only for trouble. In addition to working against virtually all the recommendations we have supplied, such values become extremely difficult to separate for purposes of analyses (but GENDER and RELIGION, coded separately are simple to combine in most software packages).

s0050 *Missing Values*

p0160 However carefully researchers plans their project, they will inevitably confront the problem of missing values. A respondent may have failed (or refused) to answer a question about his/her religion, a case may lack a clear disposition, information simply may be unavailable for a particular county, and so on. Investigators should be aware of this problem from the onset and prepare accordingly. This is so even if they plan to invoke one of the methods scholars have developed to deal with missing data because it might affect the analyses. That is because the various solutions to the problem assume that researchers treat missing data appropriately when they create the original database.

p0165 At the very least, investigators must incorporate into their codebook values to take into account the possibility of missing data—with these values distinguishing among the different circumstances under which missing information can arise. These can include “refused to answer/no answer,” “don’t know,” and “not applicable,” among others. Whatever the circumstances, researchers should assign values to them rather than simply leaving blank spaces. Simply leaving missing values blank can cause all types of logistical problems—for example, is the observation truly missing, or has the coder not yet completed it? Using an explicit missing-value code eliminates this type of confusion and can also provide information about why a specific variable is missing.

One final point is worthy of mention: Although in this entry our primary concern is with coding variables to be included in an initial and original database—not with imputing missing data (or recoding or otherwise transforming variables)—we do want to note that if the researcher has imputed missing data, she should indicate this in the final version of the codebook. The Inter-university Consortium for Political and Social Research suggests one of two possible approaches: “The first is to include two versions of any imputed variables, one being the original, including missing data codes, and the second being an imputed version, containing complete data. A second approach is to create an ‘imputation flag,’ or indicator variable, for each variable subject to imputation, set to ‘1’ if the variable is imputed and ‘0’ otherwise.” p0170

Coding Notes s0055

As we have noted earlier, the overriding goal of a codebook—and indeed the entire coding process—is to minimize the need for interpretation. As far as possible, human judgment should be removed from coding, or, when a judgment is necessary, the rules underlying the judgments should be wholly transparent to the coders and other researchers. Only by proceeding in this way can researchers help to ensure the production of reliable measures. p0175

To accomplish this in practice, analysts certainly ought be as clear as possible in delineating the values of the variables. But they also should write down a very precise set of rules for the coders (and other analysts) to follow and should include that information for each variable housed in their codebook. Such a list should be made even if investigators code the data themselves, because without it, others will not be able to replicate the research (and the measure). Along these lines, an important rule of thumb is to imagine that the researchers had to assign students the task of classifying each case by its disposition and that the only communication permitted between the researchers and the students was a written appendix to the article detailing the coding scheme. This is the way to conduct research and how it should be judged. (We do not deal with the topic of transforming original variables here, but, of course, if researchers create new variables from existing ones, they should note this.) p0180

Coding and Data Entry s0060

Once researchers have devised their codebook, they must turn to the tasks of (1) employing the codebook to assign a value for every variable for each unit under study and (2) entering these values into a statistical software program. They can perform these tasks concurrently or p0185 separately. Analysts making use of computer-assisted telephone interviewing (CATI) or computer-assisted personal interviewing (CAPI) programs, for example, do not separate the two; they us direct data entry. At the other extreme, researchers who are coding their data from a host of sources may record the assigned values on coding sheets and then transfer the information to a software program. To see why, let us return to the example of the analysts coding court DISPOSITIONS and suppose that, in addition to the values of this variable, they also desire data on the PARTY AFFILIATION of the judges deciding the case. Because she cannot obtain information on DISPOSITION and PARTY AFFILIATION from the same source, she may find it prudent to create a coding (or a transfer) sheet, assign values to each case on the variable DISPOSITION, and then do likewise for PARTY AFFILIATION. Once she has collected and coded all the data, she can enter the information (now on coding sheets) into her software package.

p0190 The rules covering coding sheets have been outlined elsewhere. What is important here is understanding the trade-off (often a necessary one) that researchers make when they enter data from sheets rather than directly into their computer. On the one hand, researchers must realize that every extra step has the potential to create error—recording information onto coding sheets and then transferring that information into a computer introduces a step that is not necessary in direct data entry. On the other hand (and even if data are derived from one source only), they should understand that coding and data entry typically represent two separate tasks—asking a singular person to perform them concurrently may also lead to errors in one, the other, or both.

p0195 Whatever choices researchers make, they should evaluate them. Reliability checks on the coding of variables are now standard—researchers drawing a random sample of cases in their study and asking someone else to recode them is a simple way to conduct them. So too analysts ought assess the reliability of the data entry process, even if they made use of sophisticated software to input the information. Although such programs may make it difficult, if not impossible, for investigators to key in wild or out-of-range values (e.g., a 7 when the only values for GENDER are 0 = male, 1 = female), they typically do not perform consistency or other checks. And when using multiple coders, it is necessary to have several coders code a set of the same observations to allow the researcher to assess reliability among the coders.

Researchers can undertake the process of cleaning their data set in several ways (e.g., running frequency distributions or generating data plots to spot outliers, or creating cross-tabulations to check for consistency across variables). But the key point is that they should do it, because as Babbie well states in a 2001 study, "'dirty' data will almost always produce misleading research findings." The same, of course, is true of data that have been collected and coded via asystematic, unthinking means—that is, via means that these recommendations are designed to thwart. p0200

See Also the Following Articles

Further Reading

Babbie, E. (2001). "The Practice of Social Research." Wadsworth, Belmont, CA.

Data Documentation Initiative. http://www.icpsr.umich.edu/DDI/ORG/index.html

Epstein, L., and King, G. (2002). The rules of inference. *University Chicago Law Rev.* **69,** 1–133.

Frankfort-Nachmias, C., and Nachmias, D. (2000). "Research Methods in the Social Sciences." Worth, New York.

Inter-university Consortium for Political and Social Research. (2002). Guide to social science data preparation and archiving. Ann, Arbor, MI. Available at: http://www.icpsr.umich.edu/ACESS/dpm.html

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithim for multiple imputation. *Am. Polit. Sci. Rev.* **95,** 49–69.

King, G., Keohane, R. O., and Verba, S. (1994). "Designing Social Inquiry: Scientific Inference in Qualitative Research." Princeton University Press, Princeton, NJ.

Little, R. J. A., and Rubin, D. B. (1987). "Statistical Analysis with Missing Data." John Wiley, New York.

Manheim, J. B., and Rich, R. C. (1995). "Empirical Political Analysis," 4th Ed. Longman, New York.

Salkind, N. J. (2000). "Exploring Research." Prentice Hall, Upper Saddle River, NJ.

Shi, L. (1997). "Health Services Research Methods." Delmar Publishers, Albany, NY.

Stark, R., and Roberts, L. (1998). "Contemporary Social Research Methods." MicroCase, Bellvue, WA.

U.S. Census Bureau. http://www.census.gov/epcd/www/naics.html

U.S. Court of Appeals Data Base. http://www.polisci.msu.edu/pljp/databases.html