Getting to know your variables: The foundation for a good working relationship with your data

Jane E. Miller Ph.D.
Rutgers University

INTRODUCTION

As experienced researchers know, each variable in a statistical analysis measures a specific concept in a particular context collected with a specific study design and data collection instrument. Because concepts, context, and study design all affect the valid range and interpretation of numeric values for those variables, it is important that students develop the habit of getting to know each of their variables before analyzing their data. Just as knowledge of the place, time, and circumstances in which someone is living or working helps them perform better in that setting, knowledge about the context pertaining to one's data can help avert blunders in data analysis. And just as someone's relationships with people are more successful if they get to know them as individuals rather than working from stereotypes, students' experiences with data will be more effective if they get acquainted with their individual variables before working with them.

Too often, however, students simply run statistics on their data without stopping to learn the substantive, real-world meaning of their variables or check the range of values. On a more advanced level, some students forge ahead and interpret of multivariate regression coefficients without stopping to think about whether a one-unit increase in an independent variable makes sense for the level of measurement and range of values in that variable (Miller 2013, chapter 10).

For more concrete illustrations of why it is a bad idea to treat all variables as if they were the same, consider the following examples of how not all variables can take on all values. A value of 10,000 makes sense in at least some contexts (places, times or groups) for annual family income in dollars, the population of a census tract, or an annual death rate per 100,000 persons. However, with rare exceptions, a value of 10,000 does *not* make sense for hourly income in dollars or birth weight in grams, and never fits number of persons in a family, a Likert scale item, a proportion, or an annual death rate per 1,000 persons. A value of –1 makes sense for temperature in degrees Fahrenheit or Celsius, *change* in rating on a 5 point scale, change in a death rate, or percentage change in income, but is completely nonsensical for temperature in degrees Kelvin, number of persons in a family, a proportion, or a death rate. Thus, it is essential that students learn not to think about variables in their analyses as generic, but instead to understand the specific concepts behind each of the variables they are studying, which will help them identify reasonable levels and ranges for each variable in their analysis.

Failing to become acquainted with one's data can lead to many types of mistakes in data preparation, model specification, and interpretation of results. This story about the experience of a young research trainee will illustrate: She came to me in the ninth week of a ten week training program, puzzled by the results of her multivariate regression. She was analyzing data from a nationally representative survey sample from a developing country circa 2002, which she had downloaded from a research data website but hadn't cleaned or evaluated before she started analyzing it. In the sample, birth weight in grams ranged up to 9,999 with a mean of 8,000 in the sample. Had she taken the time to look up the expected range of values for that concept (birth weight) and units (grams), she would have immediately seen a red flag because 9,999 grams is roughly 22 lbs., which is a typical weight for a 1 year-old, not a newborn! A second warning sign was that two-thirds of the sample had a birth weight value of 9,999 – a very high value for such a substantial share of a sample, and one that is unlikely to be explained by either outliers or

data entry errors alone. By looking at the study documentation and questionnaire, she discovered that this distribution occurred due to a skip pattern designed to minimize recall bias in birth weight reporting such that only mothers of children under age 5 years were asked about birth weight. Children aged 5 through 17 years should have been omitted from the analytic sample because the dependent variable was missing (9999) for them. If the student had familiarized herself with the concepts, units, context, and study design related to her topic and data, she could have averted the need to rerun all of her statistical analyses and rewrite major sections of her paper at the last minute to reflect the correct sample and values.

In this paper, I outline a series of steps to be conducted whenever a student or other researcher undertakes a project on a topic or data set that is new to them. Getting acquainted with variables is a multi-step process involving several resources about the data source and the topic under study. Information on attributes such as levels of measurement, range of reasonable values, skip patterns, and other missing values is essential for data preparation, including exclusion criteria for the analytic sample and creation of new variables; choice of pertinent descriptive and inferential statistics; design of correct charts and tables; and writing correct prose descriptions for the data and methods and results sections of a research paper. Results and feedback on early steps will inform later steps in the exercise. Thus, this exercise is best undertaken as part of a course or research internship in which each student is using one data set to conduct a analysis of a single research question over the course of a full semester or longer.

This "getting to know your data and variables" exercise incorporates a wide range of research methods concepts as well as univariate statistics. For students who have learned those concepts in earlier courses the exercise can be conduct over a period of two to three weeks, with feedback at intermediate steps. For students who are learning those concepts and skills in the same semester as they tackle the exercise, the steps should be spread out over a semester to ensure that they have the chance to master the definitions and ideas behind those concepts before applying them to their own data and topic.

RESOURCES NEEDED FOR THIS EXERCISE

Prior to starting this exercise, students should have committed to a specific research question so that they can identify the relevant variables in their data set and conduct focused searches to identify articles, books, web sites etc. on their topic. In addition, they will need documentation on the data source, including a description of study design, the questionnaire or other form(s) used to collect the data, and a codebook for the data set, a copy of the electronic data file, and statistical software to analyze it. Finally, they should have access to standard research methods textbook such as Chambliss and Schutt (2012) or Treiman (2009) that define and illustrate many of the concepts mentioned in the instructions for the exercise.

BACKGROUND INFORMATION ON DATA AND VARIABLES

Research question

The first step is for each student to write out their research question, including the dependent variable(s), the key independent variable(s) and any major hypothesized potential confounders, mediators, or control variables. Doing so will help them identify the full set of variables that they should include in this exercise.

Attributes of the data set

Before getting to know characteristics of the individual variables they will work with, students should spend some time learning about attributes of the data set with which

they will work. Information on these attributes will be used in later steps of this assignment when students need to identify articles or reports on their topic in a similar context so they can find external reference values of their variables to compare against observed values in their own dataset.

Restrictions on analytic sample

In many cases, students will need to impose limits on the original data set to create an analytic sample to whom their research question pertains (Chambliss and Schutt, 2012; Miller, 2013, chapter 13) , such as limiting the sample to particular demographic traits, minimum test scores, or having a specific disease. It might also mean excluding subgroups that don't meet minimum sample size, such as if there aren't enough cases in one or more subgroups of a key variable to provide sufficient statistical power and it would not be theoretically sensible to combine them with other subgroups used in the analysis. In some instances, it may be necessary to exclude cases for whom a key variable was not collected, as in the birth weight example above.

Have students make notes about these exclusions and the reasons for each, given their research question and data set. In the electronic copy of their data set, they should impose any needed restrictions and save the syntax used to make those exclusions, which they will need when verifying results and writing the data and methods sections of their research papers (Treiman, 2009).

Context of the data

The next step in getting familiar with the data is to identify the context - when, where, and to whom it pertains, also known as "the W's" (Miller, 2004, 2013). This information is critical because knowing the topic alone may not be enough to help students recognize unrealistic values of their variables. For example, suppose you are teaching an undergraduate course about global economic patterns and have assigned your students to analyze income using any data set they can find that includes a measure of annual family income variable. If one student is studying Afghanistan today, another the entire US today, another is studying the US 100 years ago, another a sample of recent GED recipients in one American city, and another is analyzing the salaries of NFL football players, they should not all expect to see the same levels and ranges of income. An annual income that would be absurdly high for a recent GED earner might be just as absurdly low for a recent NFL first round draft pick.

Much of the information about when, where, and who will come from the documentation for the data set, which should explain the study design and sampling plan. In some cases, one or more of the W's from the original data set will have been modified while creating the student's analytic sample in order to suit their research question, as noted above.

Unit of analysis

Another key step is to identify the unit of analysis, e.g., whether the student's data pertain to individual persons, families, census tracts, institutions, or some other level of aggregation (Chambliss and Schutt, 2012; Miller, 2004, 2013). Knowing unit of analysis helps ascertain plausible range of values for their variables. For example, the number of persons in a family will be much lower than the population of a census tract or a school. Information on unit of analysis should be confirmed and labeled in both the documentation and the electronic version of the data set.

Attributes of individual variables

Next, students should familiarize themselves with attributes all of the variables to be used in their analysis, much of which can be gleaned from the documentation. Some variables they will be analyzing in the same form in which they appeared in the original data set. Others might be new variables they created from variables in the original data set, such as dummy (binary) variables created from multi-category variables; categorical versions of continuous variables; aggregated variables, e.g., income calculated from several sources or scales that combine responses to multiple items; calculated variables such as body mass index computed from weight and height, or variables they have transformed by taking logarithms, standardizing, or changing scale (Miller, 2013, chapter 10). If they created new variables for their analysis, students should save the syntax so they can check for errors and modify how those variables were created, if necessary (Treiman, 2009). In those instances, they should add rows to Table 1 to show units, coding, etc., of both the original (source) variables from which the new variables were created and the new version they will use in their analysis.

Table 1 about here

Have students start by downloading[a] or creating an electronic version of Table 1, which is a grid for organizing information about the labeling, coding, units and missing value information on each of their variables. The major row headings in the table organize the variables in their analysis based on whether they are dependent variable(s), key independent variables, or control variables for their particular research question. They should also include information about variables that constitute filter questions (e.g., used to restrict their analytic sample; Chambliss and Schutt, 2012) or sampling weights used in their analysis.

Variable name and label

For each variable, students should fill in the appropriate section (row heading) of Table 1 with the *variable name* (acronym, often limited to 8 characters) used to identify the variable in their database, which they can find in the codebook for the data set and verify in the electronic database. They should also fill in the *variable label* – a longer descriptive phrase that helps convey the substantive meaning of the variable. If they rename an item in their database with a more informative variable name (e.g., "gender" instead of Q117), include the original question name in the variable label so they can track it back to the documentation and original database. Remind students that they will use the variable *label* in all tables, charts and prose descriptions of the variables and results. The variable name will rarely be referred to in the written research paper, as their readers aren't going to use their database, so there is no reason to make them flip back to the methods section to understand what a particular alphabet-soup acronym means substantively! However, it is important to retain that acronym in their notes as a reminder of what that variable is called in the dataset and codebook for the original source.

Units and categories

The next step is to fill in units and/or categories for every variable in the analysis. For all continuous variables, the pertinent information includes the system of measurement (e.g., income in dollars, Euros, or Yuan), the level of aggregation (e.g., hourly, monthly, or annual income), and the scale of measurement (e.g., income in dollars, thousands of dollars, or millions of dollars). For all nominal or ordinal variables, instead fill in category names and their associated numeric codes used in the database. For some ordinal variables such as income group or age group, units will pertain as well.

[a] The grids for Tables 1 and 2 can be downloaded from http://www.press.uchicago.edu/books/miller/multivariate/App4_10.1.pdf

For others, such as letter grade or self-rated health, units are not specified. This information can be extracted from the codebook and documentation. See Miller (2004 or 2013) or Chambliss and Schutt (2012) for more on levels of measurement and units.

Missing value codes and reasons

For all variables in Table 1, have students read the documentation and questionnaire to learn about skip patterns and other design issues that lead to some cases in their sample not having a response to one or more questions due to *valid skips* (Miller, 2013, chapter 13). In addition, they should look in the codebook for codes that identify item *non-response* – when a respondent did not answer a question that was asked of them (Chambliss and Schutt, 2012). Most datasets will designate separate numeric (or sometimes alphanumeric) codes for each of these reasons so users can distinguish among them, e.g., 97 = not applicable; 98 = module not administered to case; and 99 = item non-response. After filling in codes for each of their variables into Table 1, students should open the electronic database and update it if necessary to identify the various missing value codes as such so that those values are treated correctly during statistical analysis.

Plausible range of values

The next step is for students to fill in information on the plausible range of values for each of their variables so they can identify any out-of-range values that occur in the data. The range of credible values can be affected by definitional limits, what is conceptually plausible, and the context of measurement (Miller 2013, chapter 10). For instance, the percentage of a whole must *by definition* fall between 0 and 100; however a percentage *change* can be negative or exceed 100. Many other variables also cannot assume negative values. For example, an index constructed by summing 20 items each of which could range from 0 to 3 will have a theoretical minimum of 0 and a theoretical maximum of 60.

The *conceptually plausible range* is topic-specific, as with infant birth weight which is limited by physiological and anatomical constraints. It is also *unit-specific*: for example, live births in the United States have birth weight in grams that range from about 400 to 5,900 (Miller, 2013, table 5.4), but the corresponding range in ounces is 14 (less than 1 lb.) to 208 (13 lbs.) Finally, context (when, where and who is in a sample) will affect the range of reasonable values, as with the income examples cited earlier.

Emphasize to students that the *definitionally possible* range of values can and often does differ from the *observed* range of values in their data, which they will investigate in a later step of this exercise. For instance, although in theory a widely used depression scale (the CESD scale) could range from 0 to 60 points, in the general population the mean is between 8 and 10 and scores above 20 are rarely observed (Radloff and Locke, 1986). To help students learn about the substantively relevant range of values for each of their variables, have them to consult the published literature on their topic.

DESCRIPTIVE STATISTICS

Once students have received feedback from their instructor or research supervisor on the conceptual attributes of each variable listed in Table 1, have them complete Table 2 by filling in descriptive statistical information from their data set, the codebook for the original data source, and one or more reference sources related to each of their key variables.

Table 2 about here

Using the electronic copy of their datasets, have students run unweighted descriptive statistics on each of the variables involved in their analysis. They should then fill information from the statistical output into Table 2 on the number of cases for which

there are valid (non-missing) values of each variable, the observed minimum, maximum, and mean values for each continuous variable, and the frequency distribution and mode for each categorical variable. Remind them to save their output for constructing tables, charts, and prose descriptions for the results section of their papers, as well as for verifying their results and the steps taken to generate those results.

Frequency distribution charts

A critical next step is create a chart to display the distribution of each variable because summary statistics such as mean, median, mode and standard deviation alone can obscure important issues in the distribution of observed values (Miller, 2004, 2013, chapter 4). Consider the example of a student who used instructions from a published article to create an acculturation scale that she used as a predictor variable in a multivariate regression model. The mean of the acculturation scale was 4.56 with a standard deviation of 2.23. However, a histogram (Figure 1) revealed that the distribution was highly unusual, with three small approximately normal distributions just above values of 0, 2 and 4, gaps between those distributions, and spikes at exact values of 6.0 and 7.0.[b] As a consequence, the constructed "scale" variable was neither a continuous variable for which one-unit increases could be sensibly interpreted, nor a categorical variable with categories that could be modeled in the regression (Miller 2013, chapters 9 and 15).

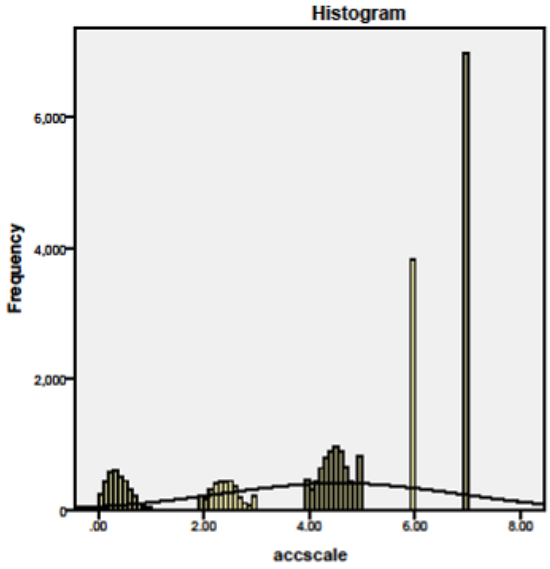

Figure 1

By examining the distribution of the scale and the original component variables, it was determined that two of the component variables (the continuous variable and one of the categorical variables) should be entered as separate predictors in the regression model in order to meet both conceptual criteria and empirical assumptions about the distributions of variables in the model. Help your student avoid analyzing "funky"

[b] Closer examination revealed that the "scale" had combined one continuous variable that was measured as a proportion with three categorical variables that each took on integer values. Scales should combine variables that share a common level of measurement and coding scheme. Moreover, the component variables were related to one another in ways that allowed for only certain values in the scale that yielded the bizarre distribution shown in Figure 1.

variables like this one by using displaying the distributions of each variable and evaluating them for plausibility.

Comparison against the codebook for the data source

Next, have your students check the distribution of values for *each* of their variables against the codebook for the data set**.** If distributions are inconsistent between those sources, have students read through the codebook and literature to identify possible reasons for discrepancies, such as units of measurement, scale (e.g., grams instead of kilograms), transformations they have made such as logged values, percentiles, or multiples of standard deviations rather than original units. Also have them consider exclusions they imposed on their sample, which might explain why such differences might actually be correct. For example, if they have restricted the sample to persons aged 25 to 34, the distribution of annual income would be expected to have a lower mean and possibly a narrower range than the distribution of annual incomes for all ages of adults from the original sample. If any distributions are inconsistent between their descriptive statistics and the information in the codebook, they should *NOT analyze the data until they have resolved the discrepancies.*

Comparison with a related population

To become familiar with which values are realistic for each of their variables, assign students to track down descriptive statistical information on those variables from the published literature for a sample that is similar to their data set in terms of location, time period and demographic characteristics (those W's again!), but is based on a different sample. Into Table 2, they should fill information from those reference sources on values against which to check plausibility of range and central tendency, or percentage distribution, as well as information about who, when, where studied.

Table 3 about here

Table 3 is an example of a comparison of a study data set (the NHANES) against national data for a similar period from Ventura et al (1999). The footnotes to Table 3 provide information about sample restrictions, definitions of variables, and calculations used to make the comparison. As in that example, if the data were collected using a complex design that involved stratification or disproportionate sampling, the statistics on the students' data should be weighted before they can be compared with national data.

Students should then compare the distribution of values for each variable in their data against the reference values from the external source of information about that variable. If they are inconsistent, have them read through the codebook and literature to identify possible reasons for discrepancies, such as different units of observation or measurement (system of measurement, level of aggregation, or scale) between their sample and the reference population. They should also watch for transformations that might explain observed differences between values observed for their sample and those from the reference population. If the values in their data set are substantially different from those used from other studies of the same concepts, students should *refrain from analyzing the data until they understand the reasons for those discrepancies.*

If students find errors based on comparison with the codebook or literature on their topic, they should correct those errors in the database and notes so they can be confident that their statistics and interpretation thereof are based on correct information. Explain that many of these steps are behind-the-scenes work that does not need to be reported in their papers, but is crucial for ensuring that the data they analyze make sense for their specific topic and data.

SUMMARY

Getting acquainted with one's data and variables is an essential step for ensuring that statistical analysis to address a research question is conceived and interpreted based on specific information about the concepts and variables in the data set at hand. Although the steps involved in this exercise are extensive and time consuming, each yields information that should be included on a paper describing the analysis, so it is time well spent. Reading the literature on the topic will provide information needed for the introduction, literature review, and discussion sections. A detailed understanding of study design and variables from documentation, questionnaire and codebook will provide information for a comprehensive data section, appropriate model specification, interpretation of statistical results, and discussion of study strengths and limitations in the concluding section of the paper. See Table 4 for a suggested timeline of conducting steps across a semester-long course.

In closing, many of the issues covered here, such as the unit of analysis, restrictions on the analytic sample, and roles of different variables in the analysis are specific to research question and data set, therefore I recommend that all researchers who are analyzing a topic that is new to them complete this "getting to know your variables" exercise.

REFERENCES

Chambliss, Daniel F., and Russell K. Schutt. 2012. M*aking Sense of the Social World: Methods of Investigation, 4th Edition*. Thousand Oaks, CA: Sage Publications.

Miller, Jane E. 2013. *The Chicago Guide to Writing about Multivariate Analysis, 2nd Edition*. Chicago: University of Chicago Press.

-----. *Supplemental Online Materials for the Chicago Guide to Writing about Multivariate Analysis, 2nd Edition*. Available online at http://www.press.uchicago.edu/books/miller/multivariate/index.html

-----. 2004. *The Chicago Guide to Writing about Numbers*. Chicago: University of Chicago Press.

Radloff, Lenore S., and Ben Z. Locke. 1986. "The Community Mental Health Assessment Survey and the CES-D Scale." In *Community Surveys of Psychiatric Disorders,* edited by M. M. Weissman, J. K. Myers, and C. E. Ross. New Brunswick, NJ: Rutgers University Press.

Treiman, Donald J. 2009. *Quantitative Data Analysis: Doing Social Research to Test Ideas*. San Francisco: Jossey-Bass.

US Department of Health and Human Services. 1997. *National Health and Nutrition Examination Survey, III, 1988–1994.* CD-ROM Series 11, no. 1. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.

Ventura, Stephanie J., Joyce A. Martin, Sally C. Curtin, and T. J. Mathews. 1999. "Births: Final Data for 1997." *National Vital Statistics Report* 47 (18). Hyattsville, MD: National Center for Health Statistics.

Table 1. Labeling, coding, and missing value information							
Variable name (e.g., acronym in your data set)	**Variable label**	**Level of measurement** (nominal, ordinal, interval, or ratio)	**Coding** (for categorical variables) **OR Units** (for continuous variables)	**Plausible range of values** (excluding missing values)	**Missing value codes** (if any)	**Skip pattern?** (e.g., conditions under which variable not collected)	**Variable from source data or created new?**
DEPENDENT VARIABLES							
INDEPENDENT VARIABLES							
Key predictor(s)							
Potential confounders or mediators							
ILLUSTRATIVE EXAMPLES							
DOCLY	Saw doctor last year	Nominal	1 = yes 2 = no	1, 2	7 = refused 8 = don't know 9 = missing	None for this variable	From source data
BWGRMS	Birth weight	Ratio	Grams	500–6000	9999 = missing	Asked only about children < 5 years old at time of survey.	From source data
CESDSCRE	Depression scale score	Ratio	Points	0–60	99 = missing	Asked only of adults	Created from items ##-##.

Table 2. Univariate statistics on variables											
	# of valid cases for that variable (excluding missing values)	**Observed values from data set[a]**				**Values & range consistent w/ codebook?**	**Reference values from external source**				**Values & range consistent w/ external source?**
Variable name (e.g. acronym on your data set)		**Min.**	**Max.**	**Mean** (for continuous variables)	**Mode**		**Min.**	**Max.**	**Mean** (for continuous variables)	**Mode**	
DEPENDENT VARIABLES											
INDEPENDENT VARIABLES											
Key predictor(s)											
Potential confounders or mediators											

[a] If the data were collected using a complex design that involved stratification or disproportionate sampling, the statistics on the students' data should be weighted before they can be compared with comparison data from the overall population from which the sample was drawn or a similar reference population (Miller, 2013 chapter 13). See Table 3 for an example.

Table 3. Birth weight, socioeconomic characteristics, and smoking behavior, NHANES III sample, 1988–1994, and all US births, 1997

	NHANES III sample[abc]	All US births,1997[d]
Birth weight		
Median (grams)	3,402	3,350
% Low birth weight (<2,500 grams)	6.8	7.5
Race/ethnicity		
Non-Hispanic white	73.4	68.4[e]
Non-Hispanic black	16.9	17.0
Mexican American	9.7	14.6
Mother's age		
% Teen mother	12.5	12.7
Mother's education		
Median (years)	12.0	12.8
% <High school	21.6	22.1
% +High school	35.0	32.4
Mother smoked while pregnant (%)	24.5	13.2
Number of cases	9,813	3,880,894

[a] Weighted to population level using weights provided with the NHANES III; sample size is unweighted.
[b] Information for NHANES III is calculated from data extracted from National Center for Health Statistics (US Department of Health and Human Services, 1997).
[c] Includes non-Hispanic white, non-Hispanic black, and Mexican American infants with complete information on family income, birth weight, maternal age, and education.
[d] Information for all US births is from Ventura et al. (1999).
[e] For consistency with the NHANES III sample, racial composition of US births is reported as a percentage of births that are non-Hispanic white, non-Hispanic black, or Mexican American, excluding births of other Hispanic origins or racial groups. When all racial/ethnic groups are considered, the racial composition is 60.1% non-Hispanic white, 15.0% non-Hispanic black, 12.9% Mexican American, 5.4% other Hispanic origin, and 6.6% other racial groups.
Adapted from Miller (2013), Table 5.5.

Table 4. Suggested order and timing of assignments for getting to know your variables			
Week #	**Step(s)**	**Readings**	**Comments**
1	1. Identify the unit of analysis 2. Write the research question and identify IV, DV, and levels of measurement for each.	Chambliss and Schutt 2012, Chapter 3 Miller, 2004, Chapter 4 Codebook for data set	
2	3. Describe the study design and context of the data set	Chambliss and Schutt 2012, Chapters 2, 4 and 5 Documentation for data set	
3	4. Identify and explain restrictions on analytic sample to fit the research question and data set		
4	5. Conduct background readings to learn about substantively plausible ranges of the independent and dependent variables 6. Impose restrictions on analytic data set in electronic copy of database. Save syntax	Literature search on topic	
5	7. Fill in Table 1, sections on units, categories, and missing values 8. Fill in missing value codes for each variable into the electronic copy of the database	Codebook for data set	
6	9. Fill descriptive statistics into Table 2 10. Fill codebook information into Table 2	Codebook for data set	Run descriptive statistics on all variables in the analysis
7	11. Fill information on outside reference source into Table 2	Literature search on topic	
8	12. Revise electronic copy of database to correct for discrepancies between data, codebook and references. Save syntax. 13. Redo descriptive statistics for any revised variables, and fill into Table 2		
9 and 10	14. Conduct statistical analysis for paper 15. Write results section of paper	Chambliss and Schutt, 2012, Chapter 8 Miller, 2004, Chapter 9	To suit research question
11	16. Write data and methods section of paper	Miller, 2004, Chapter 10	See assignment from week 2.
12	17. Write strengths and limitations for discussion section of paper	Miller, 2004, Chapters 10 and 11	