# RESEARCH METHODS 4RM - Data Collection

Author:   Ed Nelson
Department of Sociology M/S SS97
California State University, Fresno
Fresno, CA 93740
Email:  ednelson@csufresno.edu

**Note to the Instructor:** This is the fourth in a series of 13 exercises that were written for an introductory research methods class.  The first exercise focuses on the research design which is your plan of action that explains how you will try to answer your research questions.  Exercises two through four focus on sampling, measurement, and data collection.  The fifth exercise discusses hypotheses and hypothesis testing.  The last eight exercises focus on data analysis.  In these exercises we’re going to analyze data from one of the [Monitoring the Future Surveys](http://monitoringthefuture.org/) (i.e., the 2017 survey of high school seniors in the United States).  This data set is part of the collection at the Inter-university Consortium for Political and Social Research at the University of Michigan.  The data are freely available to the public and you do not have to be a member of the Consortium to use the data.  We’re going to use SDA (Survey Documentation and Analysis) to analyze the data which is an online statistical package written by the Survey Methods Program at UC Berkeley and is available without cost wherever one has an internet connection.  A weight variable is automatically applied to the data set so it better represents the population from which the sample was selected.  You have permission to use this exercise and to revise it to fit your needs.  Please send a copy of any revision to the author so I can see how people are using the exercises. Included with this exercise (as separate files) are more detailed notes to the instructors and the exercise itself.  Please contact the author for additional information.

[This page](http://ssric.org/files/Research_Methods_4RM.docx) in MS Word (.docx) format is attached.

**Goal of Exercise**

The goal of this exercise is to provide an introduction to data collection which is an integral part of any research design.  In this exercise we’re going to focus on survey research as a method of data collection.  The other elements of your research design are sampling, measurement, and data analysis which are discussed in other exercises.

**Part I—Inevitability of Error**

Error is inevitable in any research study.  It’s impossible to eliminate all sources of error.  What we do is to try to identify all sources of error and then minimize error to the extent possible.  There are a number of different types of error.  In this exercise we’ll discuss four types or sources of error in survey research – sampling, coverage, nonresponse, and measurement.

**Part II – Sampling Error**

No sample is ever a perfect representation of the population from which the sample is drawn.  Some error is always introduced when you take a sample from a population and this is called sampling error.  Imagine that your population is all high school seniors at a large urban high school that has 3,000 seniors.  We’re interested in the percent of seniors that engage in binge drinking.[[1]](%22%20%5Cl%20%22_ftn1%22%20%5Co%20%22)  We decide to select a simple random sample of 300 seniors from this population and ask each of these 300 seniors if they ever had five or more drinks in a row.[[2]](%22%20%5Cl%20%22_ftn2%22%20%5Co%20%22)  The percent of the sample that has engaged in binge drinking is our estimate of the population percent.  Now imagine selecting a second simple random sample of 300 seniors from this same population and asking them the same question.  You can immediately see that our two estimates of the percent of the population that binge drink would not be identical because the two samples would consist of different high school seniors.[[3]](%22%20%5Cl%20%22_ftn3%22%20%5Co%20%22)

Assuming that we are using probability sampling, sampling error depends on three factors.

* Size of the sample.  The larger the sample, the less the sampling error. That’s why we prefer a large sample to a small sample.  But there is a point of diminishing returns.  Once we have a sample of between 1,000 and 2,000 there isn’t much of a reduction in sampling error when we further increase the size of our sample.  For example, election polls rarely use a sample much larger than 1,500 to estimate the percent of the population that intend to vote for a particular candidate.
* Amount of variability in the population.  The more variability there is in the population, the more the sampling error.  If the entire population intended to vote for the same candidate, there wouldn’t be any sampling error.  If the population was evenly divided as to whether they were going to vote for candidate A or candidate B that would represent the maximum amount of population variability and we would have more sampling error.
* The way we select the sample.  In exercise 2RM we discussed stratification.  When we stratify our sample, sampling error decreases.  This assumes that our stratification variable is related to whatever we are trying to estimate.  (See [exercise 2RM](http://www.ssric.org/node/623) for a fuller discussion of stratification.)

**Part III – Coverage Error**

Coverage error occurs when the list of the population from which we select our sample does not perfectly match the population.  Think about the example from part 2 where we selected a sample of 300 high school seniors from the population of 3,000 seniors at a large urban high school.  We would expect the high school to have an accurate list of all seniors from which we could select our sample.  In this case the list of the population would perfectly match the population and there would be no coverage error.[[4]](%22%20%5Cl%20%22_ftn4%22%20%5Co%20%22)

Here are some examples of coverage error.

* The General Social Survey is a large national probability sample of adults in the United States conducted biannually by the National Opinion Research Center at the University of Chicago.  Prior to 2006 the sample consisted of adults living in non-institutionalized settings who spoke English.  Starting in 2006, Spanish-speaking adults were included in the sample.  While the exclusion of non-institutionalized adults and those who don’t speak English or Spanish introduces two sources of coverage error, both represent small proportions of the adult population.  The exclusion of these two groups offers both cost savings and greater ease of survey administration.  Since the coverage error is relatively small, the advantages outweigh the small increase in sampling error.  However, as more non-English and non-Spanish speaking individuals immigrate to the U.S., coverage error might increase in the future.
* The Monitoring the Future Survey of high school seniors excludes seniors in Alaska and Hawaii for both cost reasons and ease of administration.  This introduces a small amount of coverage error.  Since the survey is administered in the spring of each year, it also excludes seniors who dropped out prior to the survey administration.  This too introduces a small amount of coverage error.

These examples demonstrate that small amounts of coverage error may be tolerated for cost reasons and the ease of survey administration.  But sometimes coverage error can be quite large as you will see in the next part of this exercise.

**Part IV – Now It’s Your Turn to Consider Coverage Error**

We’re going to consider two hypothetical surveys and think about the types of coverage error that might occur.

* Our research center has been asked to do a survey of adults in our community regarding quality of life.  Specifically, we want to determine how satisfied respondents are with different areas of their local community including the local economy, level of crime, road conditions, health services, and education.  We decide to do a phone survey of households in our community.
	+ Suppose we select a sample of phone numbers from the local phone directory.  What types of coverage error would that introduce?  Do you think the amount of coverage error would be fairly small or quite large?  Why?
	+ Someone points out to us that published phone directories do not typically include cell phones so we contact a reputable research service that provides samples of both cell phone and landline numbers.  What would that do to our coverage error?  Would there be any remaining sources of coverage error?  What would they be?  Do you think they would be fairly small or quite large?  Why?
* Your research center has been asked to do a survey of all members of a particular religious group (e.g. Roman Catholics, Lutherans, Buddhists, Muslims) in a large metropolitan city (e.g., Los Angeles, Chicago, New York).  The problem is that there is no list of the population for any of these religious groups.  So you decide to do a multi-stage cluster sample.  First, you select a sample of centers of worship (e.g., churches or temples or mosques).  Then you go to each of the centers in your sample and request a membership list.[[5]](%22%20%5Cl%20%22_ftn5%22%20%5Co%20%22)  But you notice another problem.  A number of these centers have inadequate membership lists.   A close examination of these lists indicates that some former members who have died or moved away are still on the list.  To make matters worse, newer members have sometimes not been added to the list.  It’s even possible that some members appear twice on the list under different names.  What types of coverage error might these problems create?  What would be a possible solution to these problems?

**Part V – Nonresponse Error**

Nonresponse error occurs when part of your sample does not respond to your survey.  This can occur in two ways.  Sometimes we are not able to locate or contact members of the sample.  Other times sample respondents refuse to take part in the survey.  Response rates can be calculated in various ways but typically it is computed by dividing the number of respondents who complete the survey by the number of eligible respondents.  For example, if your sample consisted of 1,000 eligible individuals and 200 completed the survey, the response rate would be 20%.  If 700 completed the survey, the response rate would be 70%.

Let’s consider how nonresponse error might occur in the two examples from part 3.

* The General Social Survey reports response rates in the range of 60% to 82% with a response rate of 61% in 2016.  The GSS conducts in-person interviews although a few interviews are conducted on the phone when an in-person interview cannot be scheduled.  It’s easy to imagine that some potential respondents are never home or never available to be interviewed and that others simply refuse to be interviewed.  We also know that some types of respondents are more difficult to reach or convince to be interviewed.  Young males are especially difficult to reach particularly if we do not have their cell phone number.  Women tend to have a higher response rate than men.  Let’s say that our survey includes questions on same-sex marriage.  We know from other surveys that women are more likely to favor same-sex marriage than men.  If our sample underrepresents males, this could introduce a bias in our survey results.
* The Monitoring the Future Survey reports that recent surveys have a response rate in the 80% to 84% range.  Some students are absent on the day that the survey is administered in schools and other students refuse to participate.  Since the survey deals in part with drug use, it’s possible that students who use illegal drugs are more likely to stay home that day or refuse to participate.  This could introduce a bias in the estimate of the percent of high school seniors who use particular drugs.

**Part VI – Now It’s Your Turn to Consider Nonresponse Error**

For each of the surveys mentioned in part 4, discuss the ways that nonresponse might occur and the type of biases that such nonresponse might introduce into the survey results.

**Part VII – Measurement Error**

Measurement error occurs when the way we measure some concept (e.g., religious preference or religiosity) introduces error into the measurement process.  Imagine that a person goes into a bar to order an alcoholic drink.  The bartender asks the person how old he or she is.  If the individual is not of legal drinking age, the person might give an inaccurate answer because they have a self-interest in appearing older and obtaining an alcoholic drink.  Therefore, the bartender asks for proof of age.

Measurement error occurs in different ways.  The way we ask questions often influences what people tell us.  One of the areas that has been in the news lately has been global warming and climate change.  Research has shown that “Republicans were less likely to endorse that the phenomenon is real when it was referred to as ‘global warming’ … rather than ‘climate change’ … whereas Democrats were unaffected by question wording.”[[6]](%22%20%5Cl%20%22_ftn6%22%20%5Co%20%22)  Other research has shown that global warming is more likely than climate change to be interpreted as caused by people.[[7]](%22%20%5Cl%20%22_ftn7%22%20%5Co%20%22)

Research has also shown than question order can influence what people tell us.  In 1997 the Gallup Poll asked “Do you generally think that [Bill Clinton/Al Gore) is honest and trustworthy?”  Half of the sample (randomly selected) was asked the question with Clinton’s name first and the other half was asked with Gore’s name first.  When Clinton’s name appeared first he was much less likely to be perceived as honest and trustworthy than Gore when his name appeared first.  But when Clinton’s name appeared second there was only a small difference compared to Gore’s name appearing second.[[8]](%22%20%5Cl%20%22_ftn8%22%20%5Co%20%22)

Another way that measurement error can occur is when one of the responses is perceived as more socially desirable than the other responses.  Respondents might be reluctant to choose the less socially desirable response.  For example, cheating is generally seen as socially undesirable.  If we asked students if they had ever cheated, they might be more likely to say no because cheating is seen as socially undesirable.

**Part VIII – Now It’s Your Turn to Consider Measurement Error**

In this section we’re going to consider the Monitoring the Future Survey of high school seniors in the United States that has been conducted yearly since 1975.  There is a [website](http://monitoringthefuture.org/) that will give you a lot of information about this study.  Here’s a brief description from the website’s home page.

“Monitoring the Future is an ongoing study of the behaviors, attitudes, and values of American secondary school students, college students, and young adults. Each year, a total of approximately 50,000 8th, 10th and 12th grade students are surveyed (12th graders since 1975, and 8th and 10th graders since 1991). In addition, annual follow-up questionnaires are mailed to a sample of each graduating class for a number of years after their initial participation.”

A focus of these surveys is students’ drug use.  There were three questions asked to measure marijuana and hashish use.

* “On how many occasions (if any) have you used marijuana (grass, pot) or hashish (hash, hash oil) . . . in your lifetime?
* “On how many occasions (if any) have you used marijuana (grass, pot) or hashish (hash, hash oil) . . . during the last 12 months?
* “On how many occasions (if any) have you used marijuana (grass, pot) or hashish (hash, hash oil) . . . during the last 30 days?”

Write a paragraph discussing the types of measurement error might have occurred when students were asked these questions.

**Part IX – Conclusions**

In this exercise we have focused on surveys as our method of data collection.  Clearly there are other ways we might collect data.  Instead of asking people questions, we could observe their behavior and listen to what people say to each other.  We could also use data that others have collected.  The U.S. Census is a frequently used data set.  We could use diaries and letters as another source of data.  But regardless of the way we collect data we encounter possible sources of measurement error.

[[1]](%22%20%5Cl%20%22_ftnref1%22%20%5Co%20%22) Binge drinking is often defined as having five or more drinks in a row.

[[2]](%22%20%5Cl%20%22_ftnref2%22%20%5Co%20%22) A simple random sample is one in which every member of the population (i.e., all 3,000 seniors) has the same chance of being selected in the sample.

[[3]](%22%20%5Cl%20%22_ftnref3%22%20%5Co%20%22) Some students might be in both samples but we would expect the degree of overlap to be small.

[[4]](%22%20%5Cl%20%22_ftnref4%22%20%5Co%20%22) It’s easy to imagine that there could be some clerical error in our list which would create some coverage error but we would expect that error to be very small.

[[5]](%22%20%5Cl%20%22_ftnref5%22%20%5Co%20%22) For the purposes of this exercise ignore the confidentiality issue which would be a serious concern.

[[6]](%22%20%5Cl%20%22_ftnref6%22%20%5Co%20%22) Schuldt, J.P; S.H. Konrath, and N. Schwarz.  2011. “’Global Warming’ or ‘Climate Change’? Whether the Planet is Warming Depends on Question Wording.”  *Public Opinion Quarterly* 75, no. 1, p. 115.

[[7]](%22%20%5Cl%20%22_ftnref7%22%20%5Co%20%22) L. Whitmarsh. 2009. “What’s in a Name” Commonalities and Differences in Public Understanding of ‘Climate Change’ and ‘Global Warming.’” *Public Understanding of Science* 18, no. 4, p. 410.

[[8]](%22%20%5Cl%20%22_ftnref8%22%20%5Co%20%22) Moore, D.W. 2002. “Measuring New Types of Question-Order Effects: Additive and Subtractive.” *Public Opinion Quarterly* 66, no. 1, p. 83.