Chapter 3
Data and Its Acquisition

Once a research question has been established and some hypotheses have been derived, the next stage in the research process is to determine what type of data is needed to answer the question/hypotheses. There are two basic types of data in social science research: quantitative and qualitative. Quantitative data is data that represents items of interest numerically, and quantitative research involves examining patterns in such data using statistical methods. Examples of quantitative data include height measured in inches, IQ scores, years of schooling, earnings, counts of depressive symptoms, measures of attitudes, etc. Qualitative data represents small numbers of cases—situations, experiences, events—using data from observations, interviews, or archives that are usually not chosen using probabilistic methods. The phenomena investigated usually cannot be fully understood via quantification. For example, what is the process of death like for a dying person? How do caregivers deal with the death of a loved one who has suffered tremendously before death? What is it like to participate in an illegal activity like dog fighting? What is the life of street vendors in NYC like? Qualitative research involves examining responses to these types of questions interpretatively for common themes in order to understand human experience, often in marginal populations. A key distinction between quantitative and qualitative approaches is that much quantitative research is oriented toward making inferences about causal processes, while qualitative research is not.

Almost every broad topic can be studied qualitatively or quantitatively. For example, some research in stratification—traditionally a quantitatively-dominated area of study—has focused on the experience of being poor. A quantitative study may involve surveying a large number of people probabilistically to determine the effect of growing up in a poor neighborhood on developing childhood obesity. In contrast, a qualitative study might involve in-depth interviews with a few individuals in a poor neighborhood and ask what it is like to live in a neighborhood where there are no fresh fruits or vegetables available. Over the past decade or so, research involving *both* qualitative and quantitative data collection and analysis (called mixed methods research) has become popular. It is important, therefore, that one be able to

S.M. Lynch, *Using Statistics in Social Research: A Concise Approach*,
DOI 10.1007/978-1-4614-8573-5_3, © Springer Science+Business Media New York 2013

understand both qualitative and quantitative methods if one is to be up-to-date and well-read in one's areas of interest.

In general, the specific research question that one develops within a topic area determines the type of data and method needed to answer it. It would be impossible, for example, to investigate racial differences in obesity rates across age using a qualitative approach. Similarly, it would be impossible to fully capture the totality of the experience of becoming obese with quantitative data. Qualitative and quantitative research are often complementary: in some areas of study, it may be difficult to develop hypotheses to test quantitatively before conducting exploratory, qualitative research to understand the topic. In short, neither qualitative nor quantitative research is naturally superior to the other. Furthermore, when done well, *neither is easier to do than the other.*

3.1 Qualitative Data Acquisition

There are three broad types of qualitative research: ethnography, interviewing, and comparative-historical research, although the three overlap with each other. Ethnographers study people within their social worlds and attempt to understand and explain behavior based on the confines or rules of their world. Good ethnographic research is usually "thickly-descriptive," involves a small number of cases, and requires considerable immersion in the social world of interest. Ethnographic data is typically collected via participant-observation methods, with the degree to which the researcher becomes involved in the social world determining whether the data collection leans more toward the "participant" or "observation" end of the spectrum. On the "participation" end of the spectrum, ethnographers may fully embed themselves in the social world they are studying, like by joining a gang and engaging in their activities or living in a homeless shelter or on the street. "Observation" might be based on living in a community without taking on any formal participatory roles, but it could also be based on systematic examinations of public behavior wherever one goes. Interviewing is usually based on questions asked in face-to-face meetings. These interviews are somewhat different from those conducted by quantitative researchers, because the questions tend to be more open-ended, and key data are often quotes taken from transcriptions of the interviews.

Comparative-historical researchers attempt to understand large-scale socio-historical processes. In doing so, they often compare societies' experiences in order to reveal patterns. For example, a comparative-historical scholar may be interested in understanding why some countries more easily adopt democratic governments than others, or s/he may be interested in understanding why one country's political system evolved as it did. In order to answer such a question, a researcher may focus on historical documents or on interviewing key political figures. Some may also collect quantitative data and employ quantitative methods. Although comparative historical studies tend to be based on small numbers of cases, they tend to share more in common with statistical ways of thinking than other forms of qualitative data (Duneier 2012). This is because they tend to be interested in making claims about cause and effect. In recent years, some historical sociologists have moved in the direction of studying such topics as war, nationalism and state formation using quantitative data sets that cover the entire world over long periods of time (e.g., Wimmer 2013).

3.1.1 The "Unit of Analysis"

A key distinction between ethnographic and comparative-historical research is the *unit of analysis*. The unit of analysis is the level at which observations are made. The unit of analysis in ethnography is almost always the individual,[1] whereas the unit of analysis in comparative-historical work is always larger than the individual, e.g., states, countries, regions. As a general rule, when the unit of analysis is the individual, we call the research *micro* level research; when the unit of analysis is larger than the individual, we call the research *macro* level research. Some scholars have distinguished a third level—the "meso" level—which is intended to represent small groups or organizations. However, for our purposes, we will limit our discussion to micro and macro units.

Quantitative data may be collected at either level. Some examples of micro level quantitative data include measurements of individuals' ages, heights, weights, earnings, etc. Some examples of macro level data include state abortion rates, the proportion of the population in a state that is poor, the US unemployment rate at different points in time, etc. Notice that, while these characteristics may be aggregated from information on individuals, they are not characteristics of individuals, but of larger units, like counties, states, countries, etc.

The research question almost always determines the unit of analysis that should be investigated. For example, if we are examining whether the division of labor in society produces anomie, data must be collected at the macro level. The division of labor is a macro level phenomenon: it is not a characteristic of an individual. On the other hand, if we are examining whether poverty leads to psychological depression, the unit of analysis must be the individual.

Failing to use the appropriate unit of analysis when addressing a research question potentializes one of two fallacies: the ecological fallacy and the individualistic fallacy. The ecological fallacy is the misattribution of a relationship observed at the macro level to the micro level. The individualistic fallacy is the opposite: the misattribution of a micro level relationship to the macro level. To make these ideas concrete, consider the relationship between socioeconomic status and heart disease. Macro level studies find that richer countries have higher rates of heart disease mortality than poorer countries. Micro level studies, on the other hand, find that

[1] Often, the focus may be on the situation or conditions in which the respondent exists, but the unit at which this is examined is the individual.

	Republican	Democrat	Marginal
White	60	20	80
Nonwhite	20	0	20
Marginal	80	20	100

Table 3.1. Cross-tabulation of hypothetical political party votes by race.

	Republican	Democrat	Marginal
White	80	0	80
Nonwhite	0	20	20
Marginal	80	20	100

Table 3.2. Alternative cross-tabulation of hypothetical political party votes by race.

richer persons have lower levels of heart disease mortality than poorer persons. Thus, if we were to compare rich countries' levels of heart disease mortality to poor countries' levels of heart disease mortality and then conclude that, because the relationship between wealth and heart disease mortality is positive, richer individuals are at greater risk for heart disease than poorer individuals, we would be committing an ecological fallacy. Similarly, if we were to observe, at the individual level, that wealthier people are less likely to have heart disease than poorer persons, and conclude that wealthier countries must therefore have lower levels of heart disease than poorer countries, we would be committing an individualistic fallacy.

To demonstrate, quantitatively, a type of ecological fallacy, consider the following scenario. Suppose we know that a voting district is racially divided so that 80 % is white and 20 % is nonwhite. Suppose we also know that, in a recent election, 80 % of the district voted for a Republican candidate, while 20 % voted for a Democratic candidate. We might be tempted to conclude that whites voted for Republicans while nonwhites voted for Democrats, especially if we observed that same pattern in district after district. However, we cannot conclude this. Consider the data in Table 3.1 presented in a contingency table format—also called a cross-tabulation or "cross-tab." The table shows that nonwhites in that district were far more likely to vote Republican than whites were: 100 % of nonwhites voted Republican (20 out of 20), while 75 % of whites did (60 out of 80).

Table 3.2 presents an alternative data structure that maintains the same "marginal distributions," that is, the same proportions in the margins of the cross-tab, but the cell counts within the table differ. Under this scenario, the data do, in fact, support the conclusion we may have fallaciously jumped-to. However, as we saw, the data did not have to follow this pattern.

Technically, this ecological fallacy is of a slightly different form than the original one presented above. In fact, there are at least four forms of ecological fallacies, but their root is the same: individual inference cannot be made from aggregate data.

The data above can be "reversed" to illustrate an individualistic fallacy. Suppose we had numerous voting districts, all exactly like that shown in Table 3.1. We would observe that blacks were much more likely to vote for Republicans than whites were. We might then conclude that the proportion of voters that is black in a district is positively related to the proportion voting Republican. However, this clearly isn't the case, because all districts are 80 % white but voted Republican.

3.2 Quantitative Data Collection

Quantitative data are collected in a variety of ways. If the researcher collects the data him/herself, the data is considered "original;" if the researcher uses extant data, the data are considered "secondary." Whether the data are original or secondary to the researcher, such data are usually collected via one of three ways:

1. Face-to-face interview
2. Mail survey
3. Telephone survey

All three modes of data collection involve a written questionnaire (called an "instrument") with questions that usually have predetermined response categories. Each mode has its relative advantages and limitations, as Table 3.3 summarizes (1 is best; 3 is worst). As the table shows, mail surveys are the cheapest and the least labor intensive (one can implement a mail survey by oneself), but they are also slow and tend to produce low response rates (more on the importance of response rates later). In contrast, in-person interviews are costly, slow, and labor intensive. However, they tend to obtain the highest response rates and perhaps the most detailed data. The key advantage of the phone survey is its speed. Almost all contemporary political polls are done using phone surveys, which is why news organizations tend to be able to report opinions about Presidential speeches, debates, etc. almost immediately after they are aired.

Criterion	Mail	Phone	In-Person
Overall financial cost	1	2	3
Labor Intensity	1	2	3
Speed of data acquisition	2	1	3
Detail of information	2	3	1
Response Rate	3	2	1
Reliability and validity of responses	?	?	?

Table 3.3. Rankings of different modes of data collection along several dimensions (1 = best; 3 = worst).

In terms of reliability and validity of response—these terms to be defined more precisely later—it is difficult to determine which approach is best. For some sensitive information, the in-person interview is better than a mailed questionnaire, because a respondent may be hesitant to put his/her views in writing. On the other hand, a mailed questionnaire may be better for eliciting valid responses to some types of questions, because a respondent may be afraid to vocalize an opinion if he feels the interviewer may judge him poorly for it.

3.2.1 Sampling

Once a mode of data collection has been chosen, a researcher must decide to whom to give the questionnaire. It is generally infeasible to collect data on an entire population, and so researchers usually select a sample of individuals to interview. The importance of selecting an appropriate sample cannot be overstated. The goal of a survey is generally to provide information about a large population. Statistical theory proves that, if a sample is appropriately selected, it is possible to make very precise characterizations of the population with only a handful of sample members. However, an improperly selected sample will not provide us an accurate representation of the population; at its worst, an improperly selected sample may provide us an extremely misleading picture of the population. Furthermore, statistical theory and methods that are used to justify making inference and testing hypotheses *do not apply* to inappropriately-selected samples.

In order for a sample to represent the intended population, the sample should *ideally* be a simple random sample from the population. A simple random sample is one in which every individual in the population has equal probability of being selected.

Using a random sample may seem strange: How can randomness be a desirable quality? In fact, randomness is the only way to ensure that all characteristics of a population are represented proportionally to their existence in the population. For example, suppose we had a box containing 10 marbles of 2 colors: 5 are black, and 5 are white. Suppose I don't know how many of each color there are, but I am to estimate the proportion of each color in the box from a sample of 4 marbles. There are a total of 210 unique samples that could be drawn from this collection of 10 marbles. How did I determine the number of possible samples? The number of samples is the number of ways I can draw 4 marbles from a total of 10; by definition, this is a combination (we will discuss combinations in greater depth in the subsequent chapters). The computation for the number of combinations of x items taken from a total of n items is:

$$C(n,x) \equiv \binom{n}{x} = \frac{n!}{x!(n-x)!}, \tag{3.1}$$

# white	# black	Number (percent) of Samples
0	4	5 (2 %)
1	3	50 (24 %)
2	2	100 (48 %)
3	1	50 (24 %)
4	0	5 (2 %)

Table 3.4. Color composition of samples of four marbles taken from a population of 10 marbles, half black and half white.

where generically $k!$ is read as "k factorial" and means:

$$k \times (k-1) \times (k-2) \times \ldots \times 3 \times 2 \times 1 \times 0!, \tag{3.2}$$

and $0! = 1$ by definition.

Here, there are $\binom{10}{4} = 210$ possible samples. Of these 210 samples, other combinatorial calculations can tell us how many of these 210 samples consist of different numbers of white and black marbles. There are $\binom{5}{4} \times \binom{5}{0} = 5$ ways to draw a sample of 4 black marbles out of the 5 black marbles and no white marbles, $\binom{5}{3} \times \binom{5}{1} = 50$ ways to draw 3 black marbles and 1 white marble, $\binom{5}{2} \times \binom{5}{2} = 100$ ways to draw 2 black and 2 white marbles, $\binom{5}{1} \times \binom{5}{3} = 50$ ways to draw 1 black and 3 white marbles, and $\binom{5}{0} \times \binom{5}{4} = 5$ way to draw 4 white marbles.[2] If we were to randomly select 4 marbles—meaning each one of these 210 samples is equally likely to occur—there would be a probability of $100/210 = .48$ that we would obtain a sample in which the sample distribution was the same as the population distribution (i.e., a 50/50 % split). There would be an additional probability of $(50+50)/210 = .48$ that we would obtain a sample in which the sample distribution would be off by one marble in predicting the population distribution. Thus, with a random sample, there would be a very high probability that we would be able to make a close "guess" concerning the population distribution's composition. Indeed, there is only a probability of $(5 + 5)/210 = .04$ that we would conclude that there are either no white or no black marbles in the population. Table 3.4 summarizes the proportion of samples with composed of different numbers of black and white marbles.

We will discuss the importance of random sampling further when we begin to discuss statistics, but this example should help provide some intuition regarding why a random sample generally provides a better representation of a population than a nonrandom sample. With a nonrandom sample, there is no way to know

[2] See Exercise 5.10 and the solution; these computations constitute the basis of the hypergeometric distribution.

how much its characteristics deviate from those in the general population; with a random sample, statistical theory tells us how much the sample may vary from the population.

Most major surveys do not involve simple random sampling, but rather some variant of random sampling, like stratified sampling or cluster sampling. In stratified sampling, the population is split into two or more groups, and random sampling is conducted within each group. This approach to sampling is often used when the research is geared to examining group differences in some quantity, and the researcher needs to ensure an adequately large sample from all groups. For example, suppose I were interested in comparing Native Americans in the US to persons of all other races. I could divide the population into these two groups and then randomly select *m* persons from the Native American group and *n* persons from the "all other races" group. In this particular case, stratified sampling would be a better strategy than simple random sampling, because Native Americans constitute a tiny percentage of the population. A simple random sample, therefore, would probably not yield enough Native Americans (if any) to make the comparisons in which I may be interested.

In cluster sampling, (1) the population is broken into a number of "clusters," (2) a number of clusters are randomly selected, and then (3) everyone within each chosen cluster is selected into the sample. For example, the US could be broken down into neighborhoods, one could take a random sample of neighborhoods and then interview everyone in each selected neighborhood. Cluster sampling is often used when the population is large enough that it may be difficult or impossible to obtain a list of all members from which to sample.

Most major surveys today use a more complicated variant of cluster sampling called "multistage cluster sampling" or just "multistage sampling." In multistage cluster sampling, the population is broken into clusters at several levels (e.g., state, county, city, neighborhood, street). A random set of clusters is chosen (e.g., several states), then, within those clusters, a random set of clusters is chosen (e.g., several counties within each chosen state), and so on. At the lowest level, say households, a random set of households is chosen and then perhaps a random member of each selected household is chosen. Each of these deviations from simple random sampling requires some statistical adjustments to the data in order to make the final sample representative of the population. In this book, we will assume that we have a simple random sample; handling data from complex sample designs is beyond the scope of this introduction (see Scheaffer et al. 2012 and Lohr 1999 for details on sampling methods and adjustments that must be made).

In order to obtain a random sample, or some variant thereof, we typically need a listing of all the members in the population of interest (called a sample frame). Such a list may be difficult to obtain, if one exists at all. The multistage cluster sampling approach helps with this problem, because it breaks the population into clusters, lists for which typically do exist (like lists of states, counties within states, etc.). By the time the selection process gets to the neighborhood level, it may be quite easy to list all the houses.

Type of List	Problem
Voter registration list	only voters & persons over 18
Utility company list	only utility users
Phone book	only those with listed phone numbers
Newspaper subscription list	only subscribers
Email address list	only those with email
County property tax lists	only property owners

Table 3.5. Types of sample frames and some problems with them.

A number of strategies exist for obtaining lists of members of a population, but we must be careful to understand what potential biases may be introduced by using them. Table 3.5 lists some possible sample frames and some problems with them. For example, a common frame used in social research is a voter registration list. Voter registration lists are a matter of public record, making them easy to obtain. However, a key limitation of this frame is that it only contains registered voters and thus does not provide us with a sample that is representative of the entire population.

3.2.1.1 Some Invalid Strategies for Sampling

There are a number of strategies for sampling that are commonly used but are inappropriate for collecting data and reaching valid conclusions concerning *any* research question. Some of these include:

- Selecting friends and/or family.
- Web surveys (respondent self-selects).
- Phone-in surveys (respondent must call).
- Using students at one college only (only represents one college).
- Stopping people on the street, at the mall, at a restaurant, etc.

Some of these survey methods *might* produce valid results but will be generalizable only to a restricted population. In other words, we can only make claims about the population that is represented by our sample frame. For example, selecting students at random from a single college provides a valid way to make inference to that specific college but not about college students in general. Consumer surveys that come with a product may provide a valid means to determine what type of person purchases a product, but not about consumers in general. Furthermore, in such a case it only represents those who purchase such a product *and* care to complete the survey. A sample in which webpages are selected at random—and their owners surveyed—*may* yield a valid sample, but it only represents persons with webpages.

3.2.2 Response Rates

Response rates for surveys need to be high—usually responses rates below 70 % are unacceptable; rates above 90 % are usually considered good. The key consequences of a low response rate include (1) reduced statistical power and (2) the potential for biases. Why a low response rate reduces statistical power is easy to see. If the sample consisted of 100 potential respondents, and only 50 answered the survey, we have less information about the population from which the sample came. In fact, with such a large reduction in the sample size it may be impossible to compare some groups, because there may not be any respondents in a subsample of interest. For example, Native Americans constitute just under 1 % of the US population; thus, in a sample of 50, we would expect to have 0 respondents who were Native American.

A more subtle problem with a poor response rate is that it potentializes biases, where a "bias" is a discrepancy between a sample estimate and the true population quantity (parameter) of interest. As a simple example, suppose the goal of a survey was to determine who would win a presidential election, but no Democrats answered the survey. The bias in that case would be clear: the survey would suggest a clear win for the Republican, but only because there were no Democratic respondents. This is called "nonresponse bias."

We can sometimes tell whether results will be biased by nonresponse. If a number of sample characteristics do not match the characteristics of the population the sample is intended to represent—e.g., the age, sex, and race composition of the sample does not match that of the population—we should be wary. Sometimes determining whether biases are likely is difficult. For example, if the survey's goal were to examine some attitudes about a particularly sensitive issue (e.g., abortion attitudes, attitudes about gay marriage, etc.), and the survey makes that clear at the outset, persons who hold an unpopular perspective may be unlikely to answer the survey. If the unpopular perspective were not related to characteristics that are known at the population level, then it may be impossible to know that estimates concerning the attitude of interest will be biased.

3.2.3 Instrument and Item Construction

Once a sample is selected and a format for the survey instrument is selected (mail, phone, in-person interview), the specific questions that are to be asked—and their format and arrangement—should be determined. The design of questions is important, as they constitute the operationalization of your concepts of interest and will therefore determine whether you are able to satisfactorily answer your research question. That is, at the stage of developing a research question and deriving hypotheses, the relationships between concepts should be made clear, but the concepts themselves are still abstract until you attach particular survey questions/items to them. It is at this point that you need to be careful in designing the items in the questionnaire in order to appropriately measure your concepts.

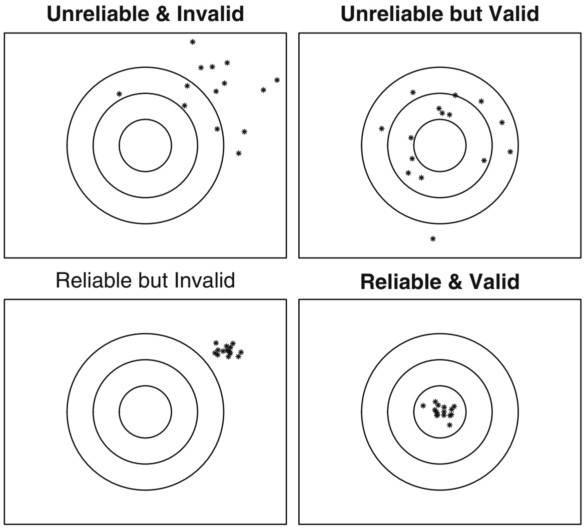

Fig. 3.1. Illustration of concepts of reliability and validity of measurement

3.2.3.1 Reliability and Validity

In the process of operationalizing your concepts via survey questions, whether you design the questions yourself or are using secondary data, you need to be sure that your measures are reliable and valid. You have much greater control over reliability and validity if you design the questions, and thus you should be careful in item construction, and ideally their location in the instrument.

Reliability can be defined as the consistency of a measure. In other words, if you ask the same question to a person again and again, or if you asked two people the same question, *all else being equal* between them, would you get the same response? If so, an item may be considered a reliable measure. Validity refers to the accuracy of a measurement, in terms of whether the question/item measures what it is intended to measure.

A classic depiction of reliability and validity is a collection of shots on a target. Figure 3.1 shows the cases graphically. If the shots are unclustered and off target, the shooter is neither reliable nor valid (upper left plot). If the shots are unclustered but centered on average over the bullseye, the shooter is valid but unreliable (upper right plot). If the shots are clustered together but away from the bullseye, the shooter is a reliable shooter (lower left plot). Finally, if the shots are clustered and centered on the bullseye, the shooter is considered reliable and valid.

Some methodologists argue that reliability is a necessary condition for validity; that is, that a measure that is unreliable cannot be valid. However, others argue, as I do here, that a measure can be a valid measure of a concept but simply not be consistent.

As an example illustrating reliability and validity, suppose we are interested in assessing individuals' health. Consider the following item:

> How would you rate your health on a scale from 1 to 100, with 1 being the worst health possible and 100 being the best?

This item seems to be a reasonable way to measure one's perceived health; however, it is not a particularly reliable measure. Why not? If I asked you this question right now, and then asked you again tomorrow, you would most likely give different answers, whether your actual health in fact varied. A number of factors, like the weather, your mood, whether you have eaten, whether you have just received news that you failed a test, etc., may affect your response, although your health hasn't actually changed. Similarly, it is quite likely that two people with exactly the same objective health will answer the item very differently. However, I would argue that, while the measure is unreliable, it is a valid measure: it seems to directly assess one's perceived health.

To make the item more reliable, we may consider reworking the possible response categories. For instance, we may change the possible responses to: Excellent, Good, Fair, or Poor. It is much more likely that, despite mood changes or other considerations, an individual will provide the same response day after day if, in fact, no real health change occurs. Similarly, it is quite likely that two individuals with the same underlying health would respond similarly. This health measure is very reliable, and studies have shown it to be a valid measure of health. In fact, this measure predicts mortality better than physician assessment and other objective health measures (Idler and Benyamini 1997)!

Now suppose we used the following as our measure of health:

> In general, how do you feel? (with Excellent, Good, Fair, or Poor as the response categories).

This item *may* be reliable, but it does not seem to be a valid measure of health. Asking respondents how they feel does not clarify whether you are referring to their current state of mental well-being, tiredness/alertness, their physical health overall, or their current general presence or lack of physical symptoms.

In general, developing reliable and valid measures is not easy, but fortunately, previous research can often guide you in developing an appropriate measure: use measures that have been validated and used before. In addition, there are at least five general rules to follow that can help in constructing good questions.

First, questions should not be double-barreled: ask ONE question at a time. Few things are more disconcerting to a respondent than being asked two questions at once—how does the respondent decide which to answer, especially if his/her responses to the two are inconsistent with one another? For example, suppose you ask "Do you favor the right of a woman to make decisions concerning her own body, or do you believe killing babies is acceptable?" Very few people, whether they are pro-choice or pro-life, would be able to answer this question. Part of the problem is that it asks two questions at once. Part of the problem is that its phrasing is vague (what does it mean for one to be able to make decisions concerning his/her own body?; what all does "killing babies" entail?); part is also that the question is leading.

Second, questions should not be leading; they should be value-neutral. The above question is a good example of an item that is leading. Certainly no one (or very few, anyway) believes that killing babies is acceptable, and so the item would not be a valid indicator of attitudes toward abortion. As another, extreme, example, suppose you ask: "Will you vote for candidate A in the next election despite the high unemployment rate during his first term (yes/no)?" It is fairly clear, given the way this question is phrased, how the interviewer would like the respondent to answer. This sort of value-laden language should not be used in an item, because it will either lead the respondent to answer in one way, or it will alienate him/her. Instead, choose neutral language that does not indicate your own predispositions.

Third, questions should not be posed as double or triple negatives. Suppose I ask: "Wouldn't you hate not having the ability to choose your own course schedule (yes or no)?" Without thinking about this item for some time, it would be difficult to decide how to answer this question, because it is posed as a either a double or triple negative: Wouldn't, hate, not. Part of the difficulty stems from the contraction at the beginning. Technically, a contraction at the beginning of a question like this implies "Would you not . . . ?" However, in common English, we generally ignore this technicality. For example, if I ask "Don't you like ice cream?," most people will respond "yes," meaning they do like ice cream. However, technically, the question asks "Do you not like ice cream," and a "yes" response implies you do not. Responses to double- and triple-negatively posed questions may therefore reflect differences in individual levels of pedantry, rather than true differences in preferences.

Fourth, make sure the question is understandable: don't use abbreviations in the question or use terminology with which the respondent may be unfamiliar. As an extension, don't ask respondents questions about issues that they will most likely not know about. If I ask "In the last 6 months have you experienced an M.I.?," many respondents will not know that I am asking whether they have had a heart attack. Substituting "myocardial infarction" for M.I. is not likely to help, either. In constructing questions, it is imperative that we ask the question so that the respondent understands what is actually being asked. As a corollary, we should not ask questions that address issues with which respondents most likely will not be familiar. For example, if I asked "Do you think Bayesian statistics is preferable to frequentist statistics?," or "Do you believe impossibility theorems are valuable," I will receive blank stares from most respondents.

Finally, ask what you intend to ask. This advice seems obvious, but it is a common source of trouble when attempting to operationalize a concept. Ultimately, this issue may have little to do with the respondent's ability to answer a question, but more to do with whether you are constructing a valid measure for the concept you are intending to measure. For example, if you are interested in the amount of money a respondent *earned* last year from work, be sure to ask the question using this terminology. Do not ask "How much money did you make last year?," because a respondent is likely to include capital gains, interest, pension distributions and other sources of income that are unearned. Or, perhaps, how much counterfeit tender s/he manufactured.

Asking a question in an appropriate fashion is only half the battle. The way in which the response categories are constructed may have just as much impact on the reliability and validity of the item as the way the question itself is asked. There are at least six rules to keep in mind when constructing answer categories.

First, response categories should be mutually-exclusive. This rule means that response categories should be non-overlapping. A respondent should not have to choose between two categories because they both contain the correct response. For example, if I ask about earnings and use ($0–$10,000, $10,000–$20,000, $20,000–$30,000, . . .) as my response categories, in which category does a person who makes $10,000 (or $20,000, or $30,000, etc.) fall? Although this rule seems obvious, especially when the outcome categories are numeric, violating this rule is commonplace. For example, the income categories in the IRS instruction book for Form 1040—the individual tax form used by virtually everyone in the US to determine annual personal income tax—has overlapping income categories.

Violations of this rule when the outcome categories are non-numeric are often more difficult to spot, but they occur often, as well. Consider, for example, the following question:

> Which one of the following would you prefer to buy at your local supermarket, if all three are available?: (A) frozen corn, (B) canned corn, or (C) fresh corn.

While it is obvious that the question seeks to determine which type of corn one prefers, the categories are not truly mutually exclusive, because one may use each type of corn, basing the decision on which to buy on its use. If my recipe calls for grilled corn, I am most likely going to buy fresh corn, because corn kernels do not fit on the grill. But, if my recipe is for a soup, I am most likely not going to buy fresh corn, because frozen or canned corn is easier to use for that purpose.

Second, the set of response categories should be exhaustive. This rule means that all possible responses should be included as choices. Although this may seem obvious, it is one of the most common problems evident in surveys. Always make sure that all the bases are covered. If you ask a question about religious affiliation, for example, be sure to include "other" and "none" as possible responses. If you ask about satisfaction with one's phonograph, be sure to include "I don't own a phonograph" as a possible response (also recall rule four above: how many know what a phonograph is these days?). More subtly, make sure that, if there is a qualitative distinction between different types of "zeros"—or similar response categories—that your response categories are capable of detecting this. For example, if I asked "When was the last time you went to church?," and my possible response categories were "more than a year ago, less than a year ago but more than a month ago, less than a month ago but more than a week ago, last week, yesterday," how would a person who has never been to church respond? Based on the response set, s/he would have to pick "more than a year ago," but there may be a distinction to be made between individuals who have never been to church and those who have, but have not been to church in more than a year. Always be aware that, if you do not provide an exhaustive set of options, the respondent may pick a category that most closely resembles the appropriate response. However, this may not be a valid response the way you intended.

Third, response categories should be meaningful. This rule simply means that the categories should be constructed so that they make relevant distinctions between individuals. For example, an income item with cutpoints every $419 is not very meaningful, and is likely to be very unreliable. Virtually no one knows what his/her income is in a given year to the dollar.

Fourth, response categories should generally not allow neutral answers, especially with difficult attitudinal items. Many argue that a "don't know" or some other neutral response category should not be an option in attitudinal questions. People generally lean one way or another when it comes to an attitudinal item, and so a question should force the respondent to decide whether s/he is more on the negative or positive side of the response set. Having a middle category may make it too easy for the respondent to not choose. This advice is debatable, because some may argue that an individual may be genuinely undecided about a particular attitudinal item and forcing a decision makes the result unreliable.

Fifth, there should not be so many response categories that a clear distinction does not exist between possible responses. The most useful data from a statistical standpoint are interval or ratio level measures. As we will discuss, items with this level of measurement facilitate a much broader set of statistical analyses than items at the nominal or ordinal level. However, there are many cases in which an item with response categories at the interval level will not be reliable (or even meaningful). For example, asking how happy an individual is with life, with "very very happy, very happy, happy, unhappy, very unhappy, very very unhappy" as the possible responses, is likely to produce considerable meaningless variance at the ends of the distribution. How does one distinguish between being very very unhappy and being only very unhappy, for instance?

Finally, response categories should produce variation. At the other extreme from providing too many response categories is not providing enough, or at least not producing response categories that create some meaningful variation. For example, if I ask a question concerning the amount of income an individual earned last year, providing (1) $0–$1,000,000 and (2) more than $1,000,000 as my response categories is not likely to produce any variation in most samples. Indeed, a "variable" is something that varies; an item that has no variation is not a variable and is useless for statistical analysis.

Overall, these rules should be followed whenever one collects his/her own data using a survey. However, these rules should also be remembered when one is considering using secondary data. The fact that survey data was collected by someone else, possibly even a reputable organization, does not mean that the items were constructed well. It may also be the case that the items were well constructed for the original purpose in which they were collected, but they may not be well constructed for your purpose. Reconsider the item above asking about one's preferred type of corn. If the survey were specifically geared toward asking about soup recipes, in particular, determining whether people prefer fresh vegetables for soup-making, then the item may be acceptable as is for that purpose. However, if the survey data were to be used by another researcher to determine individuals' general preferences for fresh vs. preserved vegetables, the item would not necessarily be reliable or valid.

3.2.3.2 Levels of Measurement

The response categories determine the *level of measurement* of the items, and so, as we will be discussing throughout the remainder of the book, they determine the type of statistical analyses that can be performed with the data. There are four basic levels of measurement: nominal, ordinal, interval, and ratio.

Nominal level response categories are unordered and unrankable. For instance, sex (male = 1; female = 2), race (white = 1; nonwhite = 2), religious affiliation (Protestant = 1, Catholic = 2, Jewish = 3, other = 4, none = 5), are all nominal level measures. The response categories cannot be ranked, and numerical values assigned to them are meaningless. They are sometimes called "qualitative" variables, because they refer to qualitative, and not quantitative, differences between people.

Ordinal level response categories are ordered and therefore rankable, but the distance between the categories is not uniform. For example, the difference between "excellent" and "good" health is not equivalent to the difference between "fair" and "poor" health, despite the fact that these four responses are ordered. Most attitudinal items use ordinal response categories, and they are often called "Likert scale" items when they have ordered categories that include words like "very" or "somewhat" to differentiate the degree of agreement or disagreement with a statement.

Interval level variables have ordered responses with equal distances between response categories. For example, temperature is an interval level measure: temperatures are ordered, and there is the same difference between 30° and 32° as there is between 60° and 62°. Technically, they both refer to the same number of calories required to move a fixed volume of water from one point to the other on the temperature scale.

Ratio level measures are interval level measures that have a true 0, and thus ratios of responses are meaningful. For example, age has a true 0. Thus, we can say that a person who is 40 is twice as old as a person who is 20. Compare this to temperature: we cannot say that 50° is twice as hot as 25°.

Although we may not be able to make every item in a questionnaire an interval level measure, we should at least be aware of the potential consequences of having nothing but nominal level measures: our ability to analyze the data—and thus the conclusions we can reach—will be more limited than they would be if we used ordinal or interval level measures. As another suggestion, consider measuring similar items at similar levels of measurement. For example, if you expect to ask a number of questions about respondents' levels of happiness and perhaps combine these items after the fact to produce a scale, they should all be measured at the same level (and ideally, with the same response categories) to facilitate their direct combination.

3.2.3.3 Guidelines for Question Placement and General Survey Strategy to Increase Response

Finally, there are some basic guidelines for making a survey more desirable, from a respondent's perspective, and hence boosting response. Additionally, there are a couple of guidelines for improving the validity and especially reliability of items, based on question placement.

First, place all demographic (and other boring) questions at the end of the survey. Demographic questions (like age, sex, race, etc.) are generally boring to respondents. Ask more interesting questions at the beginning of the survey to draw the respondent in. Then, once they have invested the time to answer these more interesting questions, they will be more likely to finish the survey.

Second, be sure to guarantee confidentiality of response, especially if some items address sensitive topics. Many respondents will not answer sensitive questions without such a guarantee, and all Institutional Review Boards (IRBs; these university committees must approve any research project before it is undertaken) will require this guarantee anyway. Note that guaranteeing confidentiality is *not* the same as guaranteeing anonymity. A guarantee of anonymity means that no one can link a respondent's name to their responses. A guarantee of confidentiality, in contrast, means that the linkage between a respondent's name and responses will exist, but only select survey administrators will have access to it. Most surveys only involve guarantees of confidentiality.

Third, be sure to discuss generally how important it is for the respondent to answer all questions in the survey, but make sure the respondent knows that participation is voluntary. If the respondent does not understand why s/he is being given the survey, s/he will be less likely to respond, especially if the questions contain sensitive information. Additionally, one requirement of research is that participation is voluntary—you cannot force individuals to participate. Explicitly saying this not only fulfills your obligation as a legitimate researcher, but it also increases the probability of response. That is, individuals are more likely to actually participate if they feel they are doing you a favor rather than being forced to respond.

Fourth, ask general questions about a topic before asking specific questions. For example, don't ask 100 questions about health conditions before asking about general health. Asking the specific questions first will make the respondent think too much about how to answer the general question and will thus make the response unreliable (and potentially invalid).

3.3 Conclusions

In this chapter, we discussed the process of survey construction and sampling, two key components to quantitative data collection in social science research. The goal of the chapter was not to provide a thorough depiction of these processes, but simply to present an overview as a prelude to the main focus of the book: the application of statistics to quantitative data collected via surveys. If you intend to collect your own data at some point, you will want to read much more detailed texts on survey methodology. However, there is nothing wrong with using secondary data. In most contemporary quantitative social science research, researchers use secondary data collected for broader purposes than for the answering of a single research question.

For example, an important social science survey is the GSS, which was mentioned in the first chapter. The GSS is a face-to-face interview conducted every 2 years (beginning in 1972) with a random sample of about 2,000 US residents in each year. The questionnaire used in the survey consists of literally hundreds of demographic, attitudinal, and other questions (some of which are the same year to year so that trends can be examined), and the data have been used in literally hundreds of published studies in many disciplines addressing many different research questions. There are literally hundreds of existing surveys like the GSS, and it is generally easier to spend some time looking for one that may suit your needs than it is to go through the process of designing and implementing your own survey.

3.4 Items for Review

Be familiar with the following concepts, terms, and items discussed in the chapter:

- Qualitative vs. quantitative data
- Unit of analysis
- Micro vs. macro level research
- Ecological fallacy
- Individualistic fallacy
- Original vs. secondary data
- Simple random sample
- Stratified sampling
- Cluster sampling
- Multistage cluster sampling
- Nonresponse bias
- Generalizability
- Reliability and validity
- Levels of measurement (nominal/ordinal/interval/ratio)
- Rules for question and answer category construction
- Confidentiality vs. anonymity
- Rules for item placement and introducing the survey to the respondent
- IRB

3.5 Homework

Answer the following questions:

1. What is wrong with the following survey question:

How old are you? (circle your response)

Under 22 22–30 30–45 45–65 65+

2. At what level of measurement is the preceding question measured?
3. At what level of measurement would be a baseball batting average?
4. What is the unit of analysis in the GSS survey described in the conclusion of the chapter?
5. What is wrong (if anything) with the following argument: Individuals with the highest levels of education have the best health. Therefore, countries with the highest levels of average educational attainment must have the best average health levels.
6. What is wrong (if anything) with the following argument: Average IQ at college A is higher than average IQ at other colleges. Thus, if I chose a student at random from college A and a student at random from some other college, the student from college A would probably have the higher IQ?
7. What is wrong (if anything) with the following argument: In the US, wealthier states tend to vote Democratic in presidential elections. Therefore, rich people are more likely to be Democrats.
8. I want to gauge support for a government-run health insurance policy. With that aim in mind, what is wrong with the following question: "Would you support a health care reform bill that does not include a government-run health insurance option? (yes/no)"
9. What type of sampling would I be using if a split the US population into 4 regions (northeast, midwest, south, and west) and then randomly selected 250 persons from each region?
10. Develop a five-item questionnaire to address the following research question: Does gender discrimination exist in the US labor market? Decide what items need to be included, and consider the reliability and validity of your items.