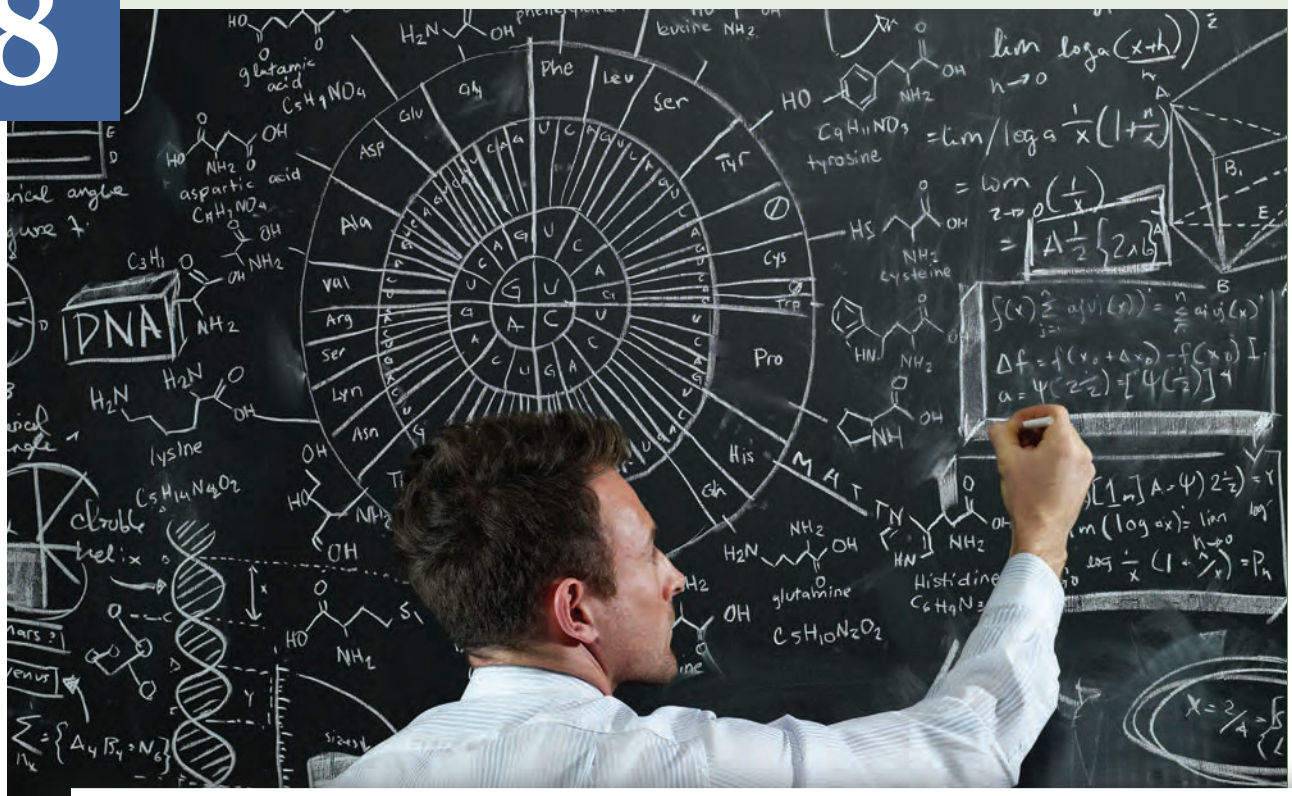
8



Foundational Research Ideas: Observation and Measurement

Learning Objectives

After reading and studying this chapter, students should be able to do the following:

- **Comprehend the importance of accurate and precise observation and measurement because these operations are at the heart of what many social scientists do.**
- **Explain how accurate measurements yield both valid and reliable scores.**
- **Differentiate between the various types of reliability and validity.**
- **Know what information is needed to make an appropriate selection of a statistic to answer questions of interest.**
- **Appreciate the various challenges and threats to collecting data, and know the steps such as pilot testing and data storage.**
- **Understand the complexity of formulating meaningful survey or questionnaire items and the challenge of survey administration and obtaining a relevant example for appropriate conclusions and generalizations.**

As a student is designing a research project, the tasks of observation and measurement are central. The fundamental goals of the social sciences are to understand and explain human behavior in all of its various forms, ranging from the individual to societal perspectives. These tasks are large endeavors for anyone interested in behavior. Sometimes these tasks seem daunting; it is hard to know where to start such a project. The foundations of research start with observation and measurement. Like building a house, if one does not have a solid foundation, whatever appears afterward will be on shaky ground. So to work toward one's goals in the social sciences, researchers acquire basic skills in observation and measurement. Presented in this chapter are key concepts to consider regarding the design of a research study.

8.1 Operational Definitions and Related Ideas

During the capstone course, students may hear about operational definitions. An **operational definition** is a translation of the key terms of the hypothesis into measurable (i.e., public, observable) quantities. If a researcher was studying depression, then he or she would need to operationally define depression in such a way as to obtain a numerical score; if a measure of hunger was desired, the score would need to be defined such that hunger is captured as a meaningful number (from a qualitative perspective, the definitions would come from nonnumeric sources). The key notion, however, is in making the connection between the behavior to be studied and the measurement of that behavior. In theory, that's where the concept of operational definition would be so crucial to what social scientists do.

So what does this all mean? The best idea would be to remember that clearly defining the key terms of research is important. Measuring human behavior, attitudes, and opinions in a meaningful way requires a rigorous approach, striving for both reliability and validity (more on these topics in the next section).

8.2 The Measurement Process

The measurement process is central to any area of research. From a research perspective, **measurement** involves how one captures the responses of individuals (either quantitatively or qualitatively) in such a manner as to allow for their systematic analysis. In any measurement process, however, there is always the possibility of error. Social scientists know this, and they keep this in mind when drawing conclusions by stating conclusions in the context of probability. Classical test theory suggests that when a measurement is obtained, that measurement (X) is composed of true score (t) plus error (e), or, $X = t + e$.

Suppose one wanted to know the height of one's best friend. That best friend has a true height—there is one correct answer. However, in measuring this best friend, there is the potential for error. The "researcher" could make an error in reading the number on a yardstick or tape measure, the friend could be wearing shoes with thick soles or could be slouching, and so on. The resulting height is composed of part true score plus part error,

which could be a result of an overestimation or an underestimation. Researchers can try to minimize error by taking other measurements and comparing results, although it should be noted that the potential for error can never be fully eliminated.

There is always a potential for error when obtaining measurements. For example, a researcher could make an error reading the correct number on a scale, but this can be avoided by conducting other measurements and comparing the results.

Similarly, measuring any aspect of a person's behavior yields a result containing true score plus error. Where does the error come from? A number of sources: For example, participants can contribute to measurement error by being unusually motivated to do well or may contribute to the measurement error by not doing their best. The instruments (surveys or questionnaires, for example) may be too demanding, too complicated, or too lengthy, leading to frustration, fatigue, or boredom. The researcher may also be a source of measurement error by being too friendly or too stern with the participants. The researcher may provide inadequate instructions about the task or may simply make errors in recording participant responses. Finally, the location and specifics of the situation may lead to measurement errors; for example, the temperature, humidity, and amount of people or space in the room may hinder the acquisition of the true score. Social scientists aim to minimize error in measurement through the use of methodology and statistics. The techniques learned throughout a student's education will help him or her make better approximations of the true score during an experiment while attempting to minimize the influence of measurement error.

Central to the observation and measurement of behavior are the concepts of reliability and validity. In other words, researchers can have confidence that behavior measures are meaningful only if they know the data—as well as the instruments used to collect that data—can be depended on for accuracy.

Reliability

Simply put, **reliability** refers to consistency in measurement. There are a number of different types of reliability, and this section provides a brief introduction to the main types.

Test-Retest Reliability

This type of reliability may perhaps be one of the easier types of reliability to understand. Test-retest reliability refers to the consistency in scores when the same test is administered twice to the same group of individuals. For example, a researcher may be interested in studying the trait of humility. Many personality traits are assumed to be relatively stable

over time, so an individual's humility levels at the beginning of the semester should not be too much different than his or her humility levels one month into the semester. Generally speaking, the longer the time passes between test and retest, the lower test-retest reliability is likely to be.

Parallel Forms/Alternate Forms Reliability

One of the benefits of the test-retest approach is that a single instrument is created and administered twice to the same group. However, one of the drawbacks to this approach is that, depending on the interval between testing, some individuals might remember some of the items from test to retest. To avoid this, someone interested in constructing a reliable test could use a parallel forms or alternative forms approach. Although related, these two approaches are technically different (Cohen & Swerdlik, 2005). In a parallel forms test, there are two versions of a test, Test A and Test B. Both Test A and Test B would be given to the same group of individuals, and the outcomes between the two test administrations could be correlated (Aiken & Groth-Marnat, 2006). With true parallel forms tests, researchers want identical means and standard deviations of test scores, but in practice, one hopes that each parallel form would correlate equivalently with other measures (Cohen & Swerdlik, 2005).

With **alternate forms** reliability, two different forms of the test are designed to be parallel, but do not meet the same criteria levels for parallel forms (e.g., nonequivalent means and standard deviations). For example, at some schools, an instructor might distribute two (or more) different versions of the test (perhaps on different colors of paper). This is usually done to minimize cheating in a large lecture hall testing situation. One hopes that the different versions of the test (that is, alternate forms) are truly equivalent. This example provides the spirit of alternate-forms testing, but does not qualify. In true alternate-forms testing, each student is asked to complete *all* alternate forms so that reliability estimates could be calculated.

Internal Consistency Reliability

Test-retest, parallel forms, and alternate forms reliability all require that a participant complete two (or more) versions of a measure. In some cases, this may not be methodologically possible or prudent. Various methods are used to estimate the reliability of a measure in a single administration rather than requiring multiple administrations or multiple forms. This gets complex, so we will not go into great detail here.

Interrater/Interobserver Reliability

Each of the preceding reliability estimates focuses on participants' responses to a test or questionnaire, attempting to address, from a particular sample, the reliability of responses. Sometimes, however, an expert panel of judges is asked to observe a particular behavior and then score that behavior based on a predetermined rating scheme. The reliability between the scores from the raters—or how much the raters agree with one another or score similarly using the same measure—is known as interrater reliability (also known as interobserver reliability, scorer reliability, or judge reliability) (Cohen & Swerdlik, 2005).

Validity and its Threats

Whereas reliability addresses consistency in measurement, **validity** addresses the question "are researchers measuring what they think they are measuring?" There are at least two major approaches to how social scientists think about validity. One approach comes from the measurement literature and how social scientists construct new measurement instruments (known as psychometrics). The classic approach here is to discuss content validity, construct validity, criterion-related validity, and face validity. The other approach comes from the study of experimental design and particular quasi-experimental designs. In fact, some refer to this latter approach as a "Cook and Campbell" approach, in part due to an influential book (1979) that brought together this conceptualization of validity, as well as a classic listing of threats to validity. Both major approaches are briefly reviewed here.

Psychometric Approach

In the classic psychometric approach, there is a trio of C's: content validity, criterion-related validity, and construct validity. All three types of validity mentioned here are important; each is necessary, but not sufficient alone, to establish validity (Morgan, Gliner, & Harmon, 2002).

Content validity refers to the composition of items that make up the test or measurement. Do the contents of the test adequately reflect the universe of ideas, behaviors, attitudes, and so forth, that comprise the behavior of interest? For example, if a researcher was interested in studying introversion and was developing an introversion inventory to measure one's level of introversion, do the actual items on the inventory capture the totality of the concept of introversion? If a student was taking the GRE subject test in psychology, content validity asks the question "are the items truly capturing your knowledge of psychology?"

Criterion-related validity refers to how the measurement outcome, or score, relates to other types of scores. A general way to think about criterion-related validity would be given that a score is now reliable, what will this score predict? Social scientists are often interested in making predictions about behavior, so criterion-related validity can be very useful in practice. Essentially, criterion-related validity addresses the predictability of current events or future events.

Construct validity has been called "umbrella validity" (Cohen & Swerdlik, 2005, p. 157) because all types of validity feed into the overall conclusion about construct validity. A construct is a hypothetical idea intended to be measured but does not exist as a tangible "thing." For example, intelligence is a construct. That is, intelligence is this hypothetical idea that we believe humans and animals possess in certain degrees. If researchers were to do a postmortem examination of a person's brain, one would not be able to extract the part of the brain known as intelligence—intelligence is not a tangible, physical entity. Intelligence is a hypothetical idea that social scientists construct, spending considerable time and energy measuring this hypothetical idea. Generally speaking, construct validity exists when a test measures what it purports to measure. Much of what is studied in the social sciences are constructs such as humility, sympathy, depression, happiness, anxiety, altruism, success, dependence, self-esteem, and so on.

Although not part of the "C" trio of validities, **face validity** is often mentioned as a type of validity, referring to whether the person taking the test believes the test measures what it purports to measure. Face validity is ascertained from the perspective of the test-taker and not from the responses to test items. If the test-taker believes that the items are unrelated to the stated purpose of the test, this might affect the quality of responses or our confidence that the test was taken seriously.

The 3 "C"s of validity (plus face validity) comprise the classic test construction approach to validity. These are important concepts to consider when developing a measure of behavior. But there are other considerations as well, such as the level of confidence we have in the conclusions we draw, or the generalizability of the results from the present study to other times, places, or settings. Cook and Campbell (1979) offered a different conceptualization of validity, and these ideas are particularly relevant.

Cook and Campbell's Approach

These authors conceptualize validity a bit differently from psychometricians when they define validity as "the best available approximation to the truth or falsity of propositions, including propositions about cause" (p. 37). According to Cook and Campbell, four different categories of validity include internal, external, statistical conclusion, and construct validity (Cook & Campbell, 1979). However, the type that is most applicable in research study design is the first type: internal validity. **Internal validity** refers to the general nature of the relationship between the independent variables and the dependent variables. The chief goal in establishing internal validity is the determination of causality: Did the manipulation of the independent variables cause changes in the dependent variables? This is a key consideration in the design of experiments and research studies. For a brief review of the types of threats that might challenge a researcher, see Table 8.1.

Table 8.1: Classic threats to internal validity

Threat to Internal Validity	Brief Definition	Research Example
History	Something happens during the experimental session that might change responses on the dependent variable.	If you are collecting data in a large classroom and the fire alarm goes off, this may impact participant's responses.
Maturation	Change in behavior can occur on its own due to the passage of time (aging), experience, etc., with the change occurring separate from the independent variable manipulation.	In a within-subjects design, you have participants view visual information on a computer screen in 200 trials. By the end of the study, participants may be fatigued and changing dependent variable responses due to time and experience.
Testing	When testing participants more than once, earlier testing may influence the outcomes of later testing.	If you design a course to help students do better on the GRE, at the beginning of the course students take the GRE, and again at the end of the course. Mere exposure to the GRE the first time may influence scores the second time, regardless of the intervention.
Instrumentation	A change occurs in the method by which you are collecting data; that is, your instrumentation changes.	If you are collecting data through a survey program on a website and the website crashes during your experiment, then you have experienced an instrumentation failure. ***(continued)***

Table 8.1: Classic threats to internal validity *(continued)*

Statistical Regression	When experimental and control group assignments are based on extreme scores in a distribution, those individuals at the extremes tend to have scores that move toward the middle of the distribution (extreme scores, when they change, become less extreme).	In a grade school setting, children who are scoring the absolute lowest on a reading ability test are given extra instruction each week on reading, and they are retested after the program is complete. Because these children's scores were at the absolute lowest part of the distribution, these scores, when they change, have nowhere else to go but up. Is that change due to the effectiveness of the reading program or statistical regression?
Selection	When people are selected to serve in different groups, such as an experimental group and a control group, are there pre-existing group differences even before the introduction of the independent variable?	In some studies, volunteers are recruited because of the type of study or potential implications (such as developing a new drug in a clinical trial). Volunteers, however, are often motivated differently than non-volunteers. This preexisting difference at the point of group selection may influence the effectiveness of the independent variable manipulation.
Mortality	Individuals drop out of a study at a differential rate in one group as compared to another group.	In your study, you start with 50 individuals in the treatment group and 50 individuals in the control group. When the study is complete, you have 48 individuals in the control group but only 32 individuals in the experimental group. There was more mortality ("loss of participants") in one group compared to the other, which means there is a potential threat to the conclusions we draw from the study.
Interaction with Selection	If some of the preceding threats happen in one group but not in the other group (selection), then these threats are said to interact with selection.	If the instrumentation fails in the control group but not in the experimental group, this is known as a selection x instrumentation threat. If something happens during the course of the study to one group but not the other, this is a selection x history threat.
Diffusion/Imitation of Treatments	If information or opportunities for the experimental group spill over into the control group, the control group can obtain some benefit of an independent variable manipulation.	In an exercise study at the campus recreation center, students in the experimental group are given specific exercises to perform in a specific sequence to maximize health benefits. Control group members also working out at the Rec Center catch on to this sequence, and start using it on their own.
Compensatory Equalization of Treatments	When participants discover their control group status and believe that the experimental group is receiving something valuable, control group members may work harder to overcome their underdog status.	In a study of academic achievement, participants in the experimental group are given special materials that help them learn the subject matter better and perform better on tests. Control group members, hearing of the advantage they did not receive, vow to work twice as hard to keep up with the experimental group and show them that they can do work at equivalent levels.
Resentful Demoralization	When participants discover their control group status and realize that the experimental group is receiving something valuable, they decide to "give up" and stop trying as hard as they normally would.	In the same academic achievement example as above, rather that vowing to overcome their underdog status, the control group simply gives up on learning the material, possibly believing that the experiment is unfair, and why should they bother to try anyway?

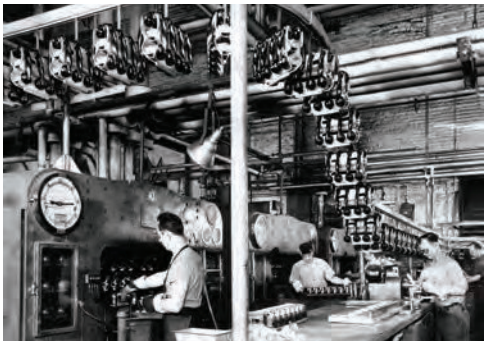
The ideas of reliability and validity are central to both the observation and measurement of behavior. These are everyday ideas that social scientists utilize to improve their work. One other note to make about the relationship between validity and reliability—an instrument can be reliable without being valid, but an instrument can only be valid when it is measured reliably. Understanding the measurement of behavior in a reliable and valid

manner is important, but what about the actual collection of data? When one relies on numerical (quantitative) scores, what do the actual numbers mean, and how might a researcher analyze the data collected? The next section covers this in detail.

Pivotal Moments in Research: The Hawthorne Studies

Generally speaking, the Hawthorne effect refers to the situation where participants in a study may band together to work harder than normal, perhaps because they have been specially selected for a study or they feel loyalty to the researchers or the experimental situation. The Hawthorne studies—so named because they were conducted at Western Electric Company's plant in Hawthorne, Illinois—began in the 1920s and ended in the 1930s. F. J. Roethlisberger of Harvard University and W. J. Dickson of the Western Electric Company were chiefly involved in these efforts, but many consultants were brought in over the course of the multiyear studies. At the time, Western Electric employed 25,000 people at the plant and served as the manufacturing and supply branch of American Telephone and Telegraph, known better today as AT&T (Baritz, 1960).

The first set of studies, beginning in November 1924, examined how different levels of lighting affected worker productivity. In one variation, individuals were tested with lighting at 10 "foot-candles." (Roughly speaking, 1 foot-candle is the amount of light that one candle generates 1 foot away from the candle.) With each successive work period, lighting decreased by 1 foot-candle. Interestingly, when lighting was decreased from 10 foot-candles to 9 foot-candles, productivity increased. In fact, productivity continued to increase with decreased lighting until about 3 foot-candles, at which point productivity decreased (Adair, 1984; Roethlisberger & Dickson, 1939). The researchers understood, if nothing else from this study, that understanding productivity was much more complicated than lighting.



Workers at the Western Electric Company factory, circa 1945.

Around April 1927, a second series of studies began, which would typically be referred to as the Relay Assembly Test Room Studies (Adair, 1984; Baritz, 1960). Experimentally speaking, Roethlisberger and Dickson became more "rigorous" in this series of studies. For example, they selected five female employees who were relay assemblers out of a large department and placed these employees in a special test room for better control of the conditions and variables to be tested. One could measure the daily and weekly output of test relays assembled by each woman as the dependent variable.

Prior to moving the female workers into the test room, researchers were able to establish a baseline of productivity based on employee records. Over the course of 270 weeks, the researchers systematically varied the conditions in the relay assembly test room, all the while recording dependent variable data on the number of test relays. For example, sometimes the amount of voluntary rest time was increased, and sometimes it was decreased. Sometimes rest breaks were increased in the morning but lengthened in the afternoon. Once workers were given Saturday mornings off (a 48-hour work week was customary at the time). Researchers also occasionally reinstated baseline control conditions. Productivity seemed to increase regardless of the manipulation introduced (Adair, 1984). In other words, even when experimental conditions were manipulated to attempt to decrease productivity, ***(continued)***

Pivotal Moments in Research: The Hawthorne Studies ***(continued)***

oftentimes productivity increased. When the employees returned to baseline control conditions, "unexpectedly, rather than dropping to pre-experiment levels, productivity was maintained" (Adair, 1984, p. 336).

There were also additional studies as part of the Hawthorne studies, such as the Mica Splitting Test Room and the Bank Wiring Room, *all with similar results*. So, what is the Hawthorne effect? Simply put, Stagner (1982) defined the Hawthorne effect as "a tendency of human beings to be influenced by special attention from others" (p. 856). In other words, when people are given special attention, they may behave differently than normal.

Questions for Critical Thinking and Reflection

- How can individuals connect the results from the Hawthorne studies to their own work history and experience? What are the motivational influences that help individuals thrive at work? Think about any sort of program or experimentation that you might believe could increase worker productivity or employee satisfaction. What types of skills do students learn at Ashford, and particularly in this capstone course, that can be applied to the workplace setting in answering these important questions?
- As a student thinks about designing a research study, how might the principles of the Hawthorne effect impact experimental design? Does the motivation of a participant in your proposed study possibly affect the validity of the data that would be collected?
- What are the conditions (that is, variables) that motivate *you* to work the way that you do? Would a higher salary lead to increases in your productivity, or a larger work space, or greater respect, or added responsibility, etc.? Fundamental attribution error is an idea from social psychology that suggests that humans attribute the behavior of others to internal decisions, while sometimes attributing their own decisions to external forces. Thus, you might think that if your coworker Joan would just work harder and not be so lazy, she would get that promotion; but when you think about yourself, you think about how the work conditions and the last manager prevented you from being promoted. Do one's thought patterns about the workplace fall in line with the fundamental attribution error? What motivates you, and how might others create an environment to help you grow and thrive?

8.3 Scales of Measurement and Statistic Selection

This chapter is about observation and measurement. When the dependent variable is measured quantitatively, researchers typically desire reliable and valid numerical scores. But how are these scores obtained? The process of translating observations into scores involves **scales of measurement**. Based in part on a seminal article by Stevens (1946), there are four general scales of measurement: nominal, ordinal, interval, and ratio. This order of presentation is important because it is generally thought that the nominal scale has the least utility in terms of value and statistical analysis options, and the ratio scale has the most utility and greater statistical options. Said another way, researchers typically prefer to have ratio scale data as compared to nominal scale data in most situations. Here is a brief overview about each scale of measurement.

Nominal

On the nominal scale, individuals are placed (or coded) into classifications or categories that are used to keep track of similarities and differences. For example, to keep track of different basketball players on the court, each wears a different number on his or her jersey. The number is used to tell the players apart on the court. Furthermore, a higher jersey number does not mean that the player is better, nor does a lower jersey number mean that the player is worse. The numbers themselves do not express relative value, but the numbers are used to track differences.

A nominal scale codes individuals into different categories, much like the number on the back of an athlete's jersey.

Numbers can also be used to track similarities. For example, if a researcher were interested in conducting a poll on campus about who people plan to vote for in the next presidential election, that researcher might also want to ask prospective voters about their political affiliation (Republican, Democrat, or Independent). As this information is recorded, this type of coding scheme might be used: 1 = Republicans, 2 = Democrats, and 3 = Independents. The numbers are used to classify people into similar categories, and different numbers are used to denote different political affiliations. Again, the numbers themselves do not have implicit meanings; that is, Independents are not one and one-half times better than Democrats, nor do Republicans have half the value of Democrats. The numbers are arbitrary placeholders allowing us to keep track of differences; these could have just as easily coded 14 = Republicans, 3 = Democrats, and 77 = Independents.

Why would a researcher want to classify nominal scale categories with numeric labels? This process facilitates data analysis in statistical programs such as SPSS (Statistical Program for the Social Sciences). But only certain types of analyses are relevant with nominal scale data. For example, in the political affiliation example, it would not make any sense to average together the 1s, 2s, and 3s; that would not yield any information. However, it would be meaningful to know that code 2 (Democrats) occurs the most frequently, making "Democrat" the most frequently observed political affiliation on that particular campus.

Ordinal

On the ordinal scale, the magnitude of the numbers mean something: In other words, a higher number means more, and a lower number means less. There is an underlying continuum expressed with the numbers on the ordinal scale. One example would be when items are rank-ordered. If the data are rank-ordered in some way, then it reflects ordinal scale numbers. So if a student were to rank order his or her top 10 movies of all time, the

number-1 movie would be the most favorite, and the number-10 movie would be the tenth favorite. In this rank-order scenario, the lower the number, the better the movie—the number has meaning.

Another assumption of the ordinal scale is that the distance or difference between adjacent numbers is not assumed to be equal; in fact, unequal intervals are assumed. When U.S. swimmer Michael Phelps won his Olympic Gold medal in 2008 in the men's 100m fly, he won with a time of 50.58 seconds. Milorad Cavic of Serbia won the silver with a time of 50.59 seconds, and Andrew Lauterstein of Australia won the bronze with a time of 51.12 seconds. The distance between the first place and second place medals was .01 second, while the distance between the second and third place medals was .13 seconds. This is what is meant by uneven intervals: The distance between first and second place is not necessarily the same distance between second place and third place.

Interval

The interval scale builds on the properties of the ordinal scale (Stevens, 1946). Like on the ordinal scale, the numbers are meaningful on the interval scale: Higher numbers mean something; there is a continuum underlying the number system. However, it differs from the ordinal scale in that the intervals are now uniform and meaningful—thus, the equal *interval* scale. One good example of the interval scale—although perhaps not readily applicable to the social sciences—is the Fahrenheit scale. A higher number means more heat, and a lower number means less heat. The intervals are uniform and meaningful: The distance between 20° and 40° is the same distance between 50° and 70°. Other examples of interval scales include latitude and longitude, altitude, a person's net financial worth, women's dress sizes, and so forth.

The Fahrenheit scale is an example of interval scale in which a lower number means less heat.

In the typical thinking about the interval scale, the number zero (0) is just another number on the scale. On the Fahrenheit scale, 0° does not mean lack of heat; it's just another number on the scale. This ends up posing a distinct challenge when scoring or quantifying data about human behavior in the social sciences. For example, what if someone were to score a 0 on an intelligence test? If 0 is just another number on the scale, what does a score of 0 mean? Is 0 a legitimate score on an intelligence test? Are negative values possible? More importantly, would a 0 on an intelligence test imply a lack of intelligence? In actual practice, there may not be many true interval scales in the social sciences; many social scientists assume that 0 is the absence of value, in effect combining it with the next type of scale, the ratio scale.

Ratio

The ratio scale shares all the characteristics of ordinal scale numbers, except for one key difference—on the ratio scale, 0 means the lack or absence of value. The ratio scale is our "usual" use of numbers: There is a quantitative dimension and an underlying continuum for the numbers used, and on the ratio scale, zero (0) is used to identify the lack of something—it is not just another number on the scale like in true interval scales. When zero is the endpoint on the scale, ratios are now meaningful. For example, 10 inches is twice as long as 5 inches (a ratio), because 0 inches means no length. Four hours is half as much as 8 hours because 0 hours means no time. Ratios are not meaningful, however, on the interval scale. Is 20° twice as warm as 10°? On a psychological test of intelligence, is someone who has an IQ of 120 twice as intelligent as someone who has an IQ of 60?

So think of ratio scale data as the usual use of numbers, such as counting the frequency of a behavior or asking a person to respond on a scale from 1 to 10. Researchers certainly make assumptions about what these numbers mean, and because the interpretation of true interval scale data is difficult for social science variables, often in practice interval and ratio scales are lumped together, referring to data as **interval / ratio**. In fact, in some places this type of data is referred to as interval/ratio scale data. In SPSS, for example, the only options for scaling variable data are nominal, ordinal, and "scale." Oftentimes, social scientists take advantage of these fuzzy boundaries between scale types. A very common scale used in survey research is a **Likert-type agreement scale**, where the items are declarative statements, and participants are asked to respond on a scale such as 1 = *strongly disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree*, and 5 = *strongly agree*. Many researchers treat these responses as interval/ratio scale data when calculating the average response for any survey particular item. However, when carefully examined, these data are not ratio scale, and probably not interval scale, but ordinal scale. Let's say that two different instructors are being evaluated at the end of the semester, and one of the items on the course evaluation is "This instructor seemed well-prepared for class." Dr. "A" might receive a mean score of 4.22, whereas Dr. "B" receives a mean score of 3.75. Researchers often treat this data like interval/ratio data, but the reported score between 1 and 5 is more similar to a rank order score. For example, is the distance on this scale between 2 and 3 the same distance between 4 and 5? Remember, on the ordinal scale, intervals do not need to be uniform, but on the interval and ratio scales, intervals must be uniform.

How would one determine what statistic to use in which data analysis situation? To answer that question, an individual needs to know: (1) on what types of scales the variables are measured, and (2) what type of conclusion is desired?

There are various ways to approach this complex issue (Morgan, Gliner, & Harmon, 2002; Vowler, 2007); a broad approach would be to ask this: Is the interest in examining the differences between groups, or in examining the associations or relationships among variables, or differences between participating groups? One would also need to know about the type of research design utilized—for example, between-groups design, within-groups design, mixed-group design, and so forth. When examining the variables (both independent and dependent), researchers need to know the scales of measurement (nominal, ordinal, interval/ratio). It is a complex process, and developing this skill set requires practice.

8.4 Data Collection Artifacts

When a data collection **artifact** occurs, the measurement process is distorted, biased, or corrupt in some fashion. It may inadvertently support the experimenter's hypothesis or detract from it. The four general categories of artifacts include physical setting, within subjects, demand characteristics, and experimenter expectancy. When designing a study, it is important for researchers to be aware of these artifacts so they can attempt to prevent them.

Physical Setting

In some cases, the physical setting may influence participant performance and lead to data artifacts. Too warm, too cool, or too humid a setting may detract from participant's true performance. Noise, general atmosphere, and crowdedness may also influence performance. If some of these conditions are out of the experimenter's control, then consistency is the goal: If room temperature is believed to affect participant performance, then an experimenter could test all participants at the same temperature. Or an experimenter could turn the potential data collection artifact into an independent variable and systematically test the hypothesis of whether or not the physical setting variable effects participants' performance. If it is thought that room temperature is affecting performance on a task, then an experimenter could test the hypothesis empirically: arranging for three different rooms, each at a different temperature, and then testing to see if temperature does indeed influence task performance.

Within Subjects

There are a number of subject-related artifacts to be aware of when collecting data. (Although human participants are no longer called "subjects"—see the case study in Chapter 2—the term is still used in some cases, such as a within-subjects design.) In particular, response sets can influence participant performance. A **response set** is a pattern of responding seen in a participant that may not accurately reflect the participant's true feelings on a topic. For example, **response set acquiescence** involves the participants getting stuck in saying "yes" repeatedly in a survey or questionnaire. If participants see their own pattern of responding as all yeses, then they may stop reading the questions carefully and answer yes to everything. The way to avoid this is to have questions worded in both directions, that is, to have both yes and no answers indicate whatever measure of interest is being studied in the experiment. See Table 8.2.

Table 8.2: Sample survey items and response set acquiescence

Susceptible to Response Set Acquiescence	Less Susceptible to Response Set Acquiescence
1. The instructor held my attention during class lectures. 2. The instructor wrote exams that fairly tested the material covered. 3. The instructor seems well prepared for class. 4. The instructor was available for extra help outside of class. 5. The instructor regularly answered students' questions.	1. The instructor was seldom able to hold my attention during class lectures. 2. The instructor wrote exams that fairly tested the material covered. 3. The instructor often appeared unprepared for class. 4. The instructor was available for extra help outside of class. 5. The instructor rarely answered students' questions.
Note: These items could be answered on a scale from 1 = *strongly disagree* to 5 = *strongly agree*.	

Response set social desirability comes from participants responding in a pattern which they believe makes them look good or look better than they are. That is, participants are presenting themselves as socially desirable when in fact they may not be. If one were to ask participants if they are racist, a researcher would probably obtain an underestimation of the actual numbers of people who could be considered racist. With socially charged issues, it is often difficult to overcome response set social desirability, but with carefully worded questions and multiple approaches to the concept (such as role-playing or simulations), such issues can be studied effectively. Clearly it would be problematic to ask a survey question "Are you a racist?" with yes/no response options because of response set social desirability. But researchers could ask multiple questions about how an individual treats or feels about individuals from different cultures and parts of the world and then start to approximate racist attitudes. Also, there are scales that are used to attempt to measure one's level of response set social desirability, such as the Marlow-Crowne Social Desirability Scale (Crowne & Marlow, 1960; Marlow & Crowne, 1961) and the lie subscale of the MMPI-2 (Pearson Education, Inc., 2007).

One other common type of within-subjects artifact is known as participants' self-perception, which occurs when the participants change themselves (on their own) during the course of the study. In many cases, the experimenter wants an assessment of current behavior (although sometimes the goal of a study is to indeed change a participant's behavior). However, this within-subjects artifact occurs when participants decide for themselves to change their own behavior, and this behavior change is not a planned part of the study. The Hawthorne study (see *Pivotal Moments in Research* earlier in the chapter) is a classic example of this.

Demand Characteristics

Demand characteristics are another data collection artifact, but rather than responding in a way to make oneself look good, demand characteristics are present when a participant is responding in a way he or she thinks the researcher wants (in a manner of speaking, giving in to the demands or expectations of the experimenter). One method of dealing with this is to disguise the nature of the study so that the participant has difficulty discerning the hypothesis and giving into the experimenter's desires. Along those lines, the participants could be uninformed about the complete nature of the study and not told about it until

its conclusion. This approach is called a **single-blind study** because the participants are "blind" (that is, they do not know) to the condition of the experiment in which they are participating (this has nothing to do with visual abilities, and this term may be considered offensive by some). One should note, however, that this involves the use of deception, and such steps should be considered at length (discussed later in Chapter 9). Often it is sufficient that the participants know in general about the study, but they do not know what specific condition or group they are in, hence not knowing how to respond to a demand characteristic. If participants cannot ascertain whether or not they are in the experimental or control group, then a single-blind study is under way and the demand characteristics can be minimized. One method to determine if the independent variable manipulation worked is simply to ask participants about it in a post-experimental interview.

Experimenter Expectancy

One additional type of data collection artifact is experimenter expectancy. This bias occurs because the experimenter (in this case, the person conducting the experimental session) accidentally influences the participants to perform in a certain, unnatural manner. This might happen if two different experimenters were used for the experimental or control groups, different instructions were used, or if one experimenter was very friendly to the experimental group but cold to the control group. To avoid these effects, the experiment could be performed in one session if feasible; experimenters can be trained to avoid experimenter expectancy cues; or a double-blind study can be performed. In a **double-blind study**, neither the participants nor the experimenter in the room know which participants are in which group (experimental or control). In this case, the experimenter cannot unknowingly provide performance cues (i.e., expectations) to the participants because the experimenter does not know which group is which. Someone else helping to administer the experiment knows of the group assignments and reveals them only when the data collection segment of the experiment is over. For more of the classic work on experimenter expectancy, see Rosenthal's work (1966, 1967).

8.5 Survey Research

A very common research methodology in the social sciences is the utilization of survey and questionnaire methods. Not only is this technique pervasive, but learning about how to better design, administer, and analyze survey outcomes is a skill that has a high chance of being used after one's undergraduate education is complete. Thus, this section provides an overview about the choices that survey researchers must answer concerning how the data are collected.

Surveys and questionnaires are often used to gain research information.

Interviews

In some ways, **in-person interviews** remain the gold standard in survey research. Interviews have fewer limitations about the types and length of survey items to be asked, and trained interviewers can use visual aids to assist during the interview (Frey & Oishi, 1995)—that is, the interviewee can see, feel, or taste a product, for example (Creative Research Systems, 2009). Interviews are thought to be one of the best ways to obtain detailed information from survey participants. The drawbacks of interviewing include high costs and the reluctance of individuals to take the amount of time to complete an interview (Creative Research Systems, 2009; Frey & Oishi, 1995). In addition to one-on-one interviews that may be prearranged, there are also intercept interviews, such as those perhaps observed at a mall where an interviewer intercepts shoppers and asks them for an interview. There are also group interviews, which some might call focus groups, where a group of people is interviewed at the same time.

Telephone

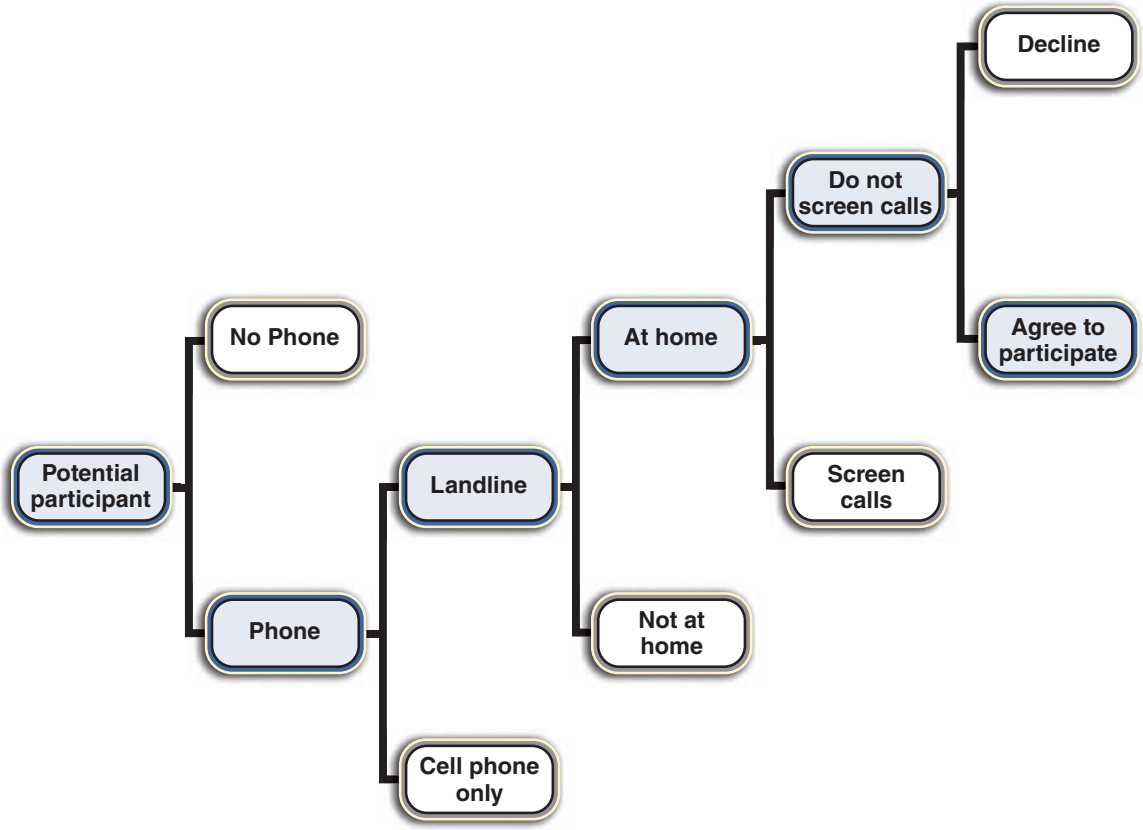
In some ways, a growing reluctance to participate in in-person interviews led to the growth of using the telephone as a modality of conducting survey research (Tuckel & O'Neill, 2002). The use of telephone methodology has increased over time but faces a number of challenges today. For instance, think about how difficult it can be to reach a willing participant on the phone—see Figure 8.1 (from Kempf & Remington, 2007).

Coverage has always been a concern of telephone research as well. That is, the greater percentage of homes with a telephone, the better the survey coverage and the better the possibility of drawing a representative sample from the population of interest. Telephone coverage in the United States has changed over time (Kempf & Remington, 2007):

- In 1920, 65% of households did not have a telephone.
- In 1970, 10% of households did not have a telephone.
- In 1986, 7–8% of households did not have a telephone.
- In 2003, less than 5% of households did not have a telephone.

Interesting trends continue to emerge with telephones, however. According to Blumberg and Luke (2010), about only 2% of U.S. households were without any telephone coverage; however, 17.5% of households have only wireless telephones. As cell-phone only households increase, so will the challenge of conducting telephone surveys.

Figure 8.1: Reaching potential participants by phone



Mail

Ever receive a survey in the mail? Was it completed by the intended person or by someone else in the household? There are advantages and disadvantages to using mailed surveys as a modality of survey data collection (de Leeuw & Hox, 2008). The advantages to mail surveys include

- relatively low cost per survey respondent—mailed surveys can be completed with a relatively small staff;
- no time pressure on the part of the survey respondent;
- the mailed survey can use visual stimuli, using different scaling techniques and visual cues for survey completions;
- the potential effect (bias) of the interviewer is removed with a mail survey;
- participants have greater privacy in responding to a mail survey; and
- if a good sample frame is available with a mailing list, the benefits of random sampling techniques can be utilized.

The potential disadvantages to mail surveys include

- potentially low response rates to mailed surveys;
- limited capabilities for complex questions, and the inability for an interviewer to clarify questions being answered on a mailed survey;
- when mail is delivered to a household, there is no guarantee that the person for whom the survey is intended is the person completing the survey; and
- the turnaround time for receiving mailed survey responses can be long.

Internet

Participating in a survey facilitated by the Internet could involve invitations through listservs, discussion groups, advertisements on search engine pages, e-mail directories, public membership directories, chat room rosters, guest lists from web pages, and of course, individual e-mail solicitations (Cho & LaRose, 1999). As a comparison to paper and pencil surveys, online/Internet surveys offer numerous advantages (Beidernikl & Kerschbaumer, 2007), including easy and inexpensive distribution to large numbers of individuals via e-mail, the participant is guided through the survey by essentially filling out a form (i.e., skip patterns are hidden from view), digital resources (e.g., video clips, sound, animation) can be incorporated into the survey design if necessary, and questions can be "required" to be answered as well as verified instantly (e.g., when asked for year of birth, if something other than a four digit number is entered, the participant can be instantly prompted to use the correct format and prevented from proceeding until making the correction).

Two key drawbacks of Internet surveys are issues of coverage and nonresponse (de Leeuw & Hox, 2008). The issue of **coverage**—that is, who has Internet access and who does not—is sometimes referred to as the digital divide (Suarez-Balcazar, Balcazar, & Taylor-Ritzler, 2009). Coverage is a problem for Internet surveys (de Leeuw & Hox, 2008), and Suarez-Balcazar et al. (2009) provided some specific examples of the possible drawbacks:

- Individuals from low-income and working class communities are less likely to have access to the Internet.
- Low-income and working-class, culturally diverse individuals are more likely to have only one computer, which would limit the potential for completing Internet-based surveys.
- Limited access often translates into limited familiarity with online/Internet applications.
- There may be cultural barriers that make Internet research more difficult to successfully accomplish (more on this in a moment).

In addition to the challenge of coverage, there is also the challenge of **representativeness**, and an Internet survey approach may not achieve the level of representativeness desired (Beidernikl & Kerschbaumer, 2007; de Leeuw & Hox, 2008). In fact, one can think about whether those replying to an Internet survey are representative of the entire population, representative of the Internet population, or even representative of a certain targeted population (Beidernikl & Kerschbaumer, 2007). Add in the complexity of culture, and one can see that well-designed Internet surveys can take a significant amount of work. Consider this example offered by Suarez-Balcazar et al. (2009):

> For instance, in the Chicago Public Schools, students speak over 100 different languages and dialects. Social scientists planning studies in these types of settings must consider how they are going to communicate with the participants' parents. Although children of first generation immigrants may be able to speak, read, and participate in Internet-based surveys in English, information such as consent forms and research protocols that are sent to the parents may need to be translated into their native language and administered using paper-and-pencil format. (p. 99)

If not used carefully, online/Internet survey researchers are capable of invading privacy (Cho & LaRose, 1999), and care should be taken to minimize the threat to invasion of privacy.

Sampling the Population

The ultimate goal of sampling the population in the design of survey research is so that a representative portion of the population can be studied (mentioned previously). Thus, by studying the sample carefully and methodically, generalizations can be drawn about the variables or behaviors of interest in the greater population. Two major types of sampling approaches exist—probability sampling and nonprobability sampling. Why sample? If the goal is to understand how the population thinks, acts, feels, believes, and so on, then why not study the entire population? First, researchers often do not have comprehensive lists of members of a population. Say, for example, a researcher wanted to survey all the citizens of Indiana. Is there a comprehensive list of all citizens available? The tax rolls might be a good start, but names and addresses are unlikely to be part of the public record. Plus, some Indiana residents may have moved, or others moved to Indiana. So it is unlikely to have an accurate and complete roster of all citizens. One can make the same generalization about the students at the college or university, all the individuals in the community with Alzheimer's disease, or all the skateboarders in town. Having an accurate roster of all the members of the population of interest would be unlikely.

By selecting a representative portion of the population to study, a researcher can make generalizations about the greater population.

In addition, there are other methodological issues as well. Because of the mathematics and probability behind sampling theory, very good samples can be drawn from populations with relatively small margins of error. Sampling is efficient. Lastly, surveying an entire population might lead to a greater number of nonrespondents, and survey researchers

become concerned about nonrespondents because if bias is driving a person's choice to not complete the survey, that may weaken the validity of the data (Dillman, Smyth, & Christian, 2009). Social scientists are better suited to select a sampling procedure that allows us to estimate any potential of sampling error in order to obtain a representative sample while minimizing bias and high nonresponse rates. Probability sampling strives to achieve each of those goals.

Probability Sampling

The overarching goal of **probability sampling** is that the sample drawn will be representative of the population if all the members of that population have an equal probability of being selected for the sample. There are various approaches of probability sampling, including simple random sampling, systematic sampling, and stratified sampling. Each is briefly described next.

The **simple random sample** is perhaps the purest form of sampling and also probably one of the rarest techniques used because of its impracticality. If the roster of the entire population were available, numbers could be assigned to all members and then selected for the sample through what is called a random number table (Babbie, 1973). Random number tables are often found at the back of statistics textbooks just for this purpose. Think of it this way: If all the names from the population were thrown into a large hat, and the researcher drew a certain target percentage for our survey, in this situation, everybody in the survey population has the same probability of being tested (Edwards & Thomas, 1993).

In a **systematic random sample**, every nth person from a list is selected (Edwards & Thomas, 1993). If at a local high school there are 2,000 students currently enrolled, and a researcher determines he or she would like to have 100 students complete the survey, each student completing the survey would have an equal chance of being selected; that is, the probability of being selected is n/N (Lohr, 2008), or in this particular example, 100/2000, or 1 out of every 20 students. So, every 20th student would be selected. After determining a random starting point (let's say #4, for example), every 20th student on the roster is selected, meaning the 4th, 24th, 44th, 64th, 84th, 104th, 124th, and so forth (Chromy, 2006).

Stratified sampling involves an approach where extra precautions are taken to ensure representativeness of the sample. Strata define groups of people who share at least one common characteristic that is relevant to the topic of the study (StatPac, 2009). For example, one might want to ensure that one's sample is representative based on gender—then one would "stratify" on gender. If the researcher knows that 55% of the population consists of females and 45% of the population consists of males, then the researcher would use a random sampling within a gender strata to extract a sample that matches the gender breakdown of the population precisely. By using relevant strata, sometimes oversampling is used to decrease sampling error from relatively small groups—that is, researchers may choose to oversample from groups less likely to respond (Edwards & Thomas, 1993). If the percentages in the population match the sample strata selected (as in the preceding gender example), this is proportionate stratification; if oversampling is used, this practice would be considered disproportionate stratification (Henry, 1990).

Nonprobability Sampling

Nonprobability methods of sampling mean just that; it is unknown what the probability is of each possible participant in the population to be selected for the study. Unfortunately, with **nonprobability sampling**, sampling error cannot be estimated (StatPac, 2009). Two key advantages to nonprobability sampling, however, are cost and convenience (StatTrek, 2009). The main approaches utilizing the nonprobability sampling approach are convenience sampling, quota sampling, snowball sampling, and a volunteer sample.

Convenience samples are just that—convenient. This might mean walking down the hallway, through a neighborhood, or other easily accessible area and handing out surveys. This technique is often used in exploratory research where a quick and inexpensive method is used to gather data (StatPac, 2009). Social scientists have long relied on convenience samples; for instance, the use of introductory-course human subject pools represent a convenience sample approach.

In **quota sampling**, the researcher also desires the strata of interest, but then recruits individuals (nonrandomly) to participate in a study (StatPac, 2009). Thus, quotas are filled with respect to the key characteristics needed for survey participants from the population.

When using the **snowball sample** technique, members of the target population of interest are asked to recruit other members of the same population to participate in the study. This procedure is often used when there is no roster of members in the population, and those members may be relatively inaccessible, such as illegal drug users, pedophiles, or members of a cult, as examples (Fife-Schaw, 2000). Snowball sampling relies on referrals and may be a relatively low-cost sampling procedure (StatPac, 2009), but there is a high probability that the individuals who participate may not be representative of the larger population.

Scaling Methods

Survey research is a complex puzzle with multiple pieces needing to be put into place before the picture is complete. Perhaps one of the most complicated parts of survey research is deciding on the **scale** by which to measure a person's attitudes, opinions, behavior, knowledge, and so forth—in fact, there are entire books on the subject (e.g., Netemeyer, Bearden, & Sharma, 2003). How a researcher shapes the possible answers can even influence the answers received. For example, Schwarz (1999) reported on some of his previous research where he had surveyed German respondents about the number of hours per week that they watch television. Two groups were asked the same question but given different response categories—these response categories are depicted in Table 8.3.

Table 8.3: How response scales can shape the results: Daily TV consumption

Low Frequency Alternatives	Percent Reporting	High Frequency Alternatives	Percent Reporting
Up to ½ hour	7.4%		
½ hour to 1 hour	17.7%		
1 hour to 1½ hours	26.5%		
1½ hours to 2 hours	14.7%		
2 hours to 2½ hours	17.7%	Up to 2½ hours	62.5%
More than 2½ hours	16.2%	2½ hours to 3 hours	23.4%
		3 hours to 3½ hours	7.8%
		3½ hours to 4 hours	4.7%
		4 to 4½ hours	1.6%
		More than 4½ hours	0.0%

Look what happens, depending on the response scale. When the scale starts low (left side of table), only 16.2% of respondents report watching more than two and one-half hours of television per day, but when the alternatives start higher on the scale (on the right side of the table), 37.5% of respondents report watching more than two and one-half hours of television per day. Just by the scale difference alone, the magnitude of this difference makes it difficult to draw meaningful conclusions. So what do researchers do about situations where surveys and scales are necessary? Social scientists rely on best practices and established research that guides the decision making necessary to select an appropriate scale.

Quick Tips for Survey Item Construction

So a researcher has determined that closed-ended items are better suited for the research needs, and this researcher is nearly ready to start generating an item pool. But before starting, it might be beneficial for him or her to think broadly for a moment about the intended measure—that broad category of human response he or she is trying to capture. Consider these broad categories offered by eSurveyPro (2009) and Rattray and Jones (2007):

- attitudes, beliefs, intentions, goals, aspirations
- knowledge, or perceptions of knowledge
- cognitions
- emotions
- behaviors and practices
- skills, or perceptions of skills
- demographics

Making decisions about which broad category (or categories) to inquire about has implications for the entire survey. For example, ask too many knowledge questions to the respondents with difficult items, and respondents may quit the survey early out of frustration. Actual skills may be difficult to capture in a survey format, but researchers may

ask respondents about their perceptions of their skills. **Demographics**, as in demographic survey questions, can be tricky as well. Ask too many demographics questions (age, gender, ethnicity, and so on), and participants may feel a sense of intrusion, and the more demographics asked, the more identifiable a participant is, even if the data are collected anonymously. Ask too few demographics and the original hypotheses may not be testable. By practicing survey skills over time, comfort levels should grow in avoiding the potential pitfalls of search research.

General advice for constructing survey items comes from many sources. See *Tips & Tools: Best Practices for Survey Item Construction* for a compilation of good ideas from multiple sources.

Tips & Tools: Best Practices for Survey Item Construction

1. Avoid double-barreled items. That is, each question should contain just one thought. A tipoff to this occurring is sometimes the use of the word *and* in a survey item.
   Don't: I like cats and dogs.
   Do: I like cats. [Next question] I like dogs.
2. Avoid using double negatives. This can cause the respondent to misread the question.
   Don't: Should the instructor not schedule an exam the same week a paper is due? (Answered from Strongly Disagree to Strongly Agree).
   Do: Should the instructor schedule an exam the same week a paper is due?
3. Avoid using implicit negatives, that is, using words like *control*, *restrict*, *forbid*, *ban*, *outlaw*, *restrain*, or *oppose*.
   Don't: Handgun use should be banned. All abortions should be outlawed.
   Do: Handgun use should be closely monitored. All abortions should be prohibited.
4. Consider offering a "no opinion" or "don't know" option.
5. To measure intensity, consider omitting the middle alternative.
   Do: Strongly disagree, disagree, ~~neutral~~, agree, and strongly agree
6. Make sure that each item is meaningful to the individuals being asked to complete the survey. That is, are the respondents competent to provide meaningful responses?
   Example to avoid: Xanax is the best prescription medication for clinical depression.
7. Use simple language, standard English as appropriate, and avoid unfamiliar or difficult words. Depending on the sample, aim for an eighth-grade reading level.
   Don't: How ingenuous are you when the professor asks if you have understood the material presented during a lecture?
   Do: Are you truthful when asked if you understood what was said in class?
8. Avoid biased questions, words, and phrases.
   Don't: Using clickers represents state-of-the-art learning technology. To what extent have clickers enhanced your learning?
   Do: Some students use clickers to answer questions. To what extent do clickers enhance learning?
9. Check to make sure your own biases are not represented in your survey items, such as leading questions.
   Don't: Do you think gas-guzzling SUVs are healthy for the environment?
   Do: Do you think SUVs are healthy for the environment?
10. Do not get more personal than necessary to adequately address your hypotheses. Focus on "need to know" items and not "nice to know" items (helps control for survey length). ***(continued)***

Tips & Tools: Best Practices for Survey Item Construction ***(continued)***

11. Try to be as concrete as possible; items should be clear and free from ambiguity. Avoid using acronyms or abbreviations that are not widely understood.
    Example to avoid: The *DSM-IV-TR* is a more accurate diagnostic tool for PSTD patients than the ICD-10.
12. Start the survey with clear instructions, and make sure the first few questions are nonthreatening. Typically, present demographic questions at the end of the survey. If asking too many demographic items, respondents may be concerned that their responses are not truly anonymous.
13. If the response scales change within a survey, include brief instructions about this so that respondents will be more likely to notice the change.
14. If your survey is long, be sure to put the most important questions first—in a long survey, respondents may become fatigued or bored by the end of the survey.
15. Be sure to frame questions in such a way as to minimize response set acquiescence. Ask reverse-scored questions (that is, strongly disagreeing is a positive outcome).
    Example: This course is a waste of time. A positive answer would be Strongly Disagree.

Sources: Babbie (1973), Cardinal (2002), Converse and Presser (1986), Crawford and Christensen (1995), Edwards and Thomas (1993), eSurveyPro (2009), Fink and Kosecoff (1985), HR-Survey (2008), Jackson (1970), McGreevy (2008), and University of Texas at Austin (2007).

Chapter Summary

The foundational principles of the scientific approach in the social sciences—observation and measurement—are presented in this chapter. In designing an experiment or quasi-experiment, basic fundamental decisions have to be made concerning independent and dependent variables. For dependent variables, how will they be measured, and if measured quantitatively, on what scale will they be measured? What operations will be followed to ensure reliability and validity of the data gathered and the conclusions drawn? Once the foundational questions are answered, then a plethora of practical matters must be considered, such as avoiding confounding variables, avoid data collection artifacts (and threats to validity), pilot testing, manipulation checks, and data collection and storage. A greater understanding of the survey research approach can be an applicable skill after completing an undergraduate education. Developing scale items measured appropriately with proper statistical analyses and conclusions drawn are abilities that social scientists can use in a variety of situations and are not limited to any particular content area or research question of interest.

Questions for Critical Thinking

- Think about the perceptions you had about social sciences before you began your formal, college-level study? Did you think that sociology would be "easy" compared to some of the other disciplines you might have studied? How do you think about sociology now? Is it as easy as you once thought? What components of an education in sociology are you finding the most worthwhile, and which components seem disconnected from other avenues of study you are pursuing?

- At Ashford you have completed a number of courses in different disciplines, and probably other courses in the social sciences outside of sociology (courses in psychology, criminal justice, anthropology, economics, just to name a few of the possibilities). How does a sociological approach to studying human behavior differ from the approaches of other social sciences in studying human behavior? To what extent are these principles of observation and measurement similar or different to the approaches in other social science disciplines?

Concept Check

1. When researchers are interested in the consistency of measurements, they are interested in
   a. reliability.
   b. validity.
   c. artifactuality.
   d. kurtosis.

2. On the _____ scale of measurement, observations, behaviors, scores, or individuals are placed into classifications or categories.
   a. interval
   b. nominal
   c. ratio
   d. ordinal

3. Response set acquiescence describes a condition where participants in a study repeatedly
   a. say "I don't know."
   b. do not show up for the study.
   c. say "yes."
   d. say "I don't care."

4. With regard to a survey research approach, the concept of coverage depicts
   a. how many questions are asked about each topic.
   b. how many days per week telephone surveyors work.
   c. the amount of topics covered in any one particular survey.
   d. the degree of access to the survey technology used.

5. Which of the following is an example of nonprobability sampling?
   a. stratified sample
   b. snowball sample
   c. simple random sample
   d. systematic random sample

Answers: 1) a, 2) b, 3) c, 4) d, 5) b

Web Links

This website describes different types of validity with additional examples of each: http://writing.colostate.edu/guides/research/relval/pop2b.cfm

This website describes the four different scales of measurement with additional examples plus links for more information about scales of measurement: http://stattrek.com/ap-statistics-1/measurement-scales.aspx

This website describes some real-life examples of double-blind studies and helps to demonstrate their importance and helpfulness in experimental research: http://www.consumerhealth.org/articles/display.cfm?ID=19991119144210

This website provides additional examples and definitions of different types of sampling, with links to additional resources: http://www.stat.yale.edu/Courses/1997-98/101/sample.htm

This website describes in more detail Likert-type scales and some of the key points in their use: http://intelligentmeasurement.wordpress.com/2007/11/20/likert-scale-surveys-best-practices/

Key Terms

alternate forms A type of reliability where two different formats of the test are designed to be highly similar to one another, but alternate forms reliability does not meet the same criteria levels for parallel forms.

artifact A distortion in the measurement process where the outcomes are biased or corrupted.

construct validity When a test measures what it purports to measure. Also known as umbrella validity.

content validity The determination as to whether or not the composition of items that make up a test measure the universe of ideas, behaviors, and attitudes that comprise the behavior of interest.

convenience sampling The sampling practice often used in exploratory research where a quick and inexpensive method is used to gather data by gathering participants who are conveniently available for the purposes of data collection.

coverage The issue of participant access to the survey technology being used.

criterion-related validity The assessment of how the measurement outcome, or score, relates to other types of scores.

demographics These are survey questions that inquire about subject variable characteristics; asking too many demographic questions may make a respondent identifiable, even in an anonymous survey.

double-blind study When neither the study participants nor the experimenter are aware of the conditions being administered during the course of an experiment in order to prevent bias.

face validity The assessment of whether or not the person taking the test believes that the test is measuring what it purports to measure.

in-person interviews A research methodology that allows an interviewer and a participant to build rapport through conversation and eye contact, which might allow for asking deeper questions about the topic of interest. This presents fewer limitations about the types and length of survey items to be asked.

internal validity Represents the confidence that the scores being measured truly represent the underlying concepts.

interval/ratio data An interval scale presents numbers in a meaningful way and provides equal intervals including "0." In a ratio scale, numbers are used in the typical fashion, where 0 = a lack of something. The two scales of measurement are typically combined in research since their interpretation individually can present challenges.

Likert-type agreement scale A survey response scale that has a five-point scale, measuring from one pole of disagreement to the other pole of agreement with each of the scale points having a specific verbal description.

measurement How the responses of individuals are captured for the purposes of research.

nonprobability sampling The sampling practice where the probability of each participant being selected for a study is unknown, and sampling error cannot be estimated.

operational definition A concise definition that exhibits precisely what is being measured.

probability sampling The sampling practice where the probability of each participant being selected for a study is known, and sampling error can be estimated.

quota sampling The sampling practice where a researcher identifies a target population of interest and then recruits individuals (non-randomly) of that population to participate in a study.

reliability Refers to consistency in measurement.

representativeness The assumption that a sample will resemble all qualities of the general population in order to ensure that results of a sample can be applied to the whole general population.

response set A pattern of responding seen in a participant that may not accurately reflect the participant's true feelings on a topic.

response set acquiescence When participants get stuck in the trend of responding "yes" repeatedly in a survey or questionnaire.

response set social desirability When participants respond in a pattern they believe makes them look good or look better than they are.

scale A tool used to measure the attitudes, perceptions, behaviors, and so forth of a person chosen to best represent a study.

scales of measurement Tools used to translate observations into scores; includes nominal, ordinal, interval, and ratio scales.

simple random sample This sampling practice is the purest form of sampling and also probably one of the rarest techniques used where everybody in the survey population has the same probability of being tested.

single-blind study A study in which the participants do not know if they are part of the experimental group or the control group.

snowball sample The informal procedure where the researcher makes an initial round of contacts to solicit participants for a study but then invites those contacts to invite others to participate.

stratified sampling The practice of dividing a sample into subcategories (strata) in a way that identifies existing subgroups (such as gender) in a general population order to make a sample the same proportion as displayed in a population.

systematic random sample The sampling practice in which every *n*th person from a sample is selected.

validity The determination as to whether or not researchers are truly "measuring what they think they are measuring" for the purposes of their research.