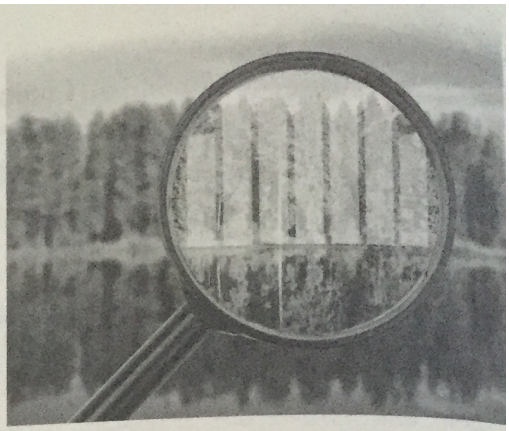
CHAPTER 3



Measurements, Mistakes, and Misunderstandings

Thought Questions

1. Suppose you were interested in finding out what people felt to be the most important problem facing society today. Do you think it would be better to give them a fixed set of choices from which they must choose or an open-ended question that allowed them to specify whatever they wished? What would be the advantages and disadvantages of each approach?
2. You and a friend are each doing a survey to see if there is a relationship between height and happiness. Without discussing in advance how you will do so, you both attempt to measure the height and happiness of the same 100 people. Are you more likely to agree on your measurement of height or on your measurement of happiness? Explain, discussing how you would measure each characteristic.
3. A newsletter distributed by a politician to his constituents gave the results of a "nationwide survey on Americans' attitudes about a variety of educational issues." One of the questions asked was, "Should your legislature adopt a policy to assist children in failing schools to opt out of that school and attend an alternative school—public, private, or parochial—of the parents' choosing?" From the wording of this question, can you speculate on what answer was desired? Explain.
4. You are at a swimming pool with a friend and become curious about the width of the pool. Your friend has a 12-inch ruler, with which he sets about measuring the width. He reports that the width is 15.771 feet. Do you believe the pool is exactly that width? What is the problem? (Note that .771 feet is 9 1/4 inches.)
5. If you were to have your intelligence, or IQ, measured twice using a standard IQ test, do you think it would be exactly the same both times? What factors might account for any changes?

3.1 Simple Measures Don't Exist

In the last chapter, we listed Seven Critical Components that need to be considered when someone conducts a study. You saw that many decisions need to be made, and many potential problems can arise when you try to use data to answer a question. One of the hardest decisions is contained in Component 4—that is, in deciding exactly what to measure or what questions to ask. In this chapter, we focus on problems with defining measurements and on the subsequent misunderstandings and mistakes that can result. When you read the results of a study, it is important that you understand exactly how the information was collected and what was measured or asked. Consider something as apparently simple as trying to measure your own height. Try it a few times and see if you get the measurement to within a quarter of an inch from one time to the next. Now imagine trying to measure something much more complex, such as the amount of fat in someone's diet or the degree of happiness in someone's life. Researchers routinely attempt to measure these kinds of factors.

3.2 It's All in the Wording

You may be surprised at how much answers to questions can change based on simple changes in wording. Here are two examples.

EXAMPLE 3.1 **You Get What You Ask for: Two Polls on Immigration**

In the spring of 2013, the question of immigration into the United States was at the forefront of many political discussions. So, polling organizations naturally wanted to know how the public felt about increasing versus decreasing the number of legal immigrants allowed into the United States. Just ask, right? Well, it's not quite so simple. Here are two polls conducted within a few days of each other, with similar wording:

- A Fox News poll conducted April 20–22, 2013, asked, *"Do you think the United States should increase or decrease the number of LEGAL immigrants allowed to move to this country?"*
- A CBS News/New York Times poll conducted April 24–28, 2013, asked, *"Should LEGAL immigration into the United States be kept at its present level, increased, or decreased?"*

The results of the two polls are shown in the following table.

Poll	Increase	Decrease	Stay Same	Unsure
Fox News	28%	55%	10%	7%
CBS/NYTimes	25%	31%	35%	8%

While the percentage responding in favor of increasing legal immigration is fairly similar in the two polls, the percentage in favor of decreasing it or keeping it the same are much different. Why? A close look at the wording reveals the answer. When the option "kept at its present level" was given in the question (CBS/New York Times Poll), 35% of respondents chose it. But when that option was not given and respondents had to come up with it on their own (Fox News Poll), only 10% did so. The lesson is that small changes in wording can make a big difference in how people respond in surveys. You should always find out exactly what was asked when reading the results of a survey. (*Source*: http://pollingreport.com/immigration.htm, accessed May 11, 2013.) ■

EXAMPLE 3.2 **Is Marijuana Easy to Buy but Hard to Get?**

Refer to the detailed report on the companion website labeled as Original Source 13: "2003 CASA National Survey of American Attitudes on Substance Abuse VIII: Teens and Parents," which describes a survey of teens and drug use. One of the questions (number 36, p. 44) asked teens about the relative ease of getting cigarettes, beer, and marijuana. About half of the teens were asked about "buying" these items and the other half about "obtaining" them. The questions and percent giving each response were:

"Which is easiest for someone of your age to buy: cigarettes, beer or marijuana?"
"Which is easiest for someone of your age to obtain: cigarettes, beer or marijuana?"

Response	Version with "buy"	Version with "obtain"
Cigarettes	35%	39%
Beer	18%	27%
Marijuana	34%	19%
The Same	4%	5%
Don't know/no response	9%	10%

Notice that the responses indicate that beer is easier to "obtain" than is marijuana, but marijuana is easier to "buy" than beer. The subtle difference in wording reflects a very important difference in real life. Regulations and oversight authorities have made it difficult for teenagers to buy alcohol, but not to obtain it in other ways. ■

Many pitfalls can be encountered when asking questions in a survey or experiment. Here are some of them; each will be discussed in turn:

1. Deliberate bias
2. Unintentional bias
3. Desire to please
4. Asking the uninformed
5. Unnecessary complexity
6. Ordering of questions
7. Confidentiality versus anonymity

Deliberate Bias

Sometimes, if a survey is being conducted to support a certain cause, questions are deliberately worded in a biased manner. Be careful about survey questions that begin with phrases like "Do you agree that. . . ." Most people want to be agreeable and will be inclined to answer "yes" unless they have strong feelings the other way. For example, suppose an anti-abortion group and a pro-choice group each wanted to conduct a survey in which they would find the best possible agreement with their position. Here are two questions that would each produce an estimate of the proportion of people who think abortion should be completely illegal. Each question is almost certain to produce a different estimate:

1. Do you agree that abortion, the murder of innocent beings, should be outlawed?
2. Do you agree that there are circumstances under which abortion should be legal, to protect the rights of the mother?

Appropriate wording should not indicate a desired answer. For instance, a Gallup Poll conducted in May 2013 (Saad, 2013) contained the question "Do you think abortion should be legal under any circumstances, legal only under certain circumstances, or illegal in all circumstances?" Notice that the question does not indicate which answer is preferable. In case you're curious, 26% of respondents thought it should always be legal, 20% thought it should always be illegal, and 52% thought it depends on the circumstance. (The remaining 2% had no opinion.).

Unintentional Bias

Sometimes questions are worded in such a way that the meaning is misinterpreted by a large percentage of the respondents. For example, if you were to ask people what drugs they use, you would need to specify if you mean prescription drugs, illegal drugs, over-the-counter drugs, or common substances such as caffeine. If you were to ask people to recall the most important date in their life, you would need to clarify if you meant the most important calendar date or the most important social engagement with a potential partner. (It is unlikely that anyone would mistake the question as being about the shriveled fruit, but you can see that the same word can have multiple meanings.)

Desire to Please

Most survey respondents have a desire to please the person who is asking the question. They tend to understate their responses about undesirable social habits and opinions, and vice versa. For example, in recent years estimates of the prevalence of cigarette smoking based on surveys do not match those based on cigarette sales. Either people are not being completely truthful or lots of cigarettes are ending up in the garbage. Political pollsters, who are interested in surveying only those who will actually vote, learned long ago that it is useless to simply ask people if they plan to vote. Most of them will say yes, because that's the socially correct answer to give. Instead, the pollsters ask questions to establish a history of voting, such as "Where did you go to vote in the last election?"

Asking the Uninformed

People do not like to admit that they don't know what you are talking about when you ask them a question. In a paper on the topic, Graeff (2007, p. 682) summarizes much of the research by noting that "survey respondents have freely given opinions about fictitious governmental agencies, congressional bills, nonexistent nationalities, fictitious political figures, and have even given directions to places that do not exist." The following example illustrates some ways that survey researchers have tried to reduce this problem, but you will see that they have not been completely successful.

EXAMPLE 3.3 Giving Opinions on Fictional Brands

To test the extent to which people don't want to admit they don't know something, Graeff (2007) asked students at his university to provide numerical ratings of their opinion of six brands of running shoes. Included in the list were five real brands and a fictional brand named Pontrey. Different versions of the survey were given to different groups of students. One version provided the option of "Don't Know/No Opinion," and in that version, only 18% of the respondents expressed an opinion about Pontrey shoes. But another version provided no such option, and in that case, 88% of respondents gave an opinion about the fictional Pontrey shoes. They did not want to admit that they had never heard of them. (Not surprisingly, the average rating for Pontrey shoes was lower than the average for any of the real brands.) In other versions of the survey, students were first asked to rate their knowledge of each brand on a scale from 1 (not familiar at all) to 7 (very familiar) before giving their opinion of the brand. Admitting that they had little knowledge of Pontrey shoes reduced the uninformed responses somewhat, but not as much as offering the "don't know" option. ■

Unnecessary Complexity

If questions are to be understood, they must be kept simple. A question such as "Shouldn't former drug dealers not be allowed to work in hospitals after they are released from prison?" is sure to lead to confusion. Does a yes answer mean they should or they shouldn't be allowed to work in hospitals? It would take a few readings to figure that out.

Another way in which a question can be unnecessarily complex is to actually ask more than one question at once. An example would be a question such as "Do you support the president's health care plan because it would ensure that all Americans receive health coverage?" If you agree with the idea that all Americans should receive health coverage, but disagree with the remainder of the plan, do you answer yes or no? Or what if you support the president's plan, but not for that reason?

Ordering of Questions

If one question requires respondents to think about something that they may not have otherwise considered, then the order in which questions are presented can change the results. For example, suppose a survey were to ask, "To what extent do you think

teenagers today worry about peer pressure related to drinking alcohol?" and then ask, "Name the top five pressures you think face teenagers today." It is quite likely that respondents would use the idea they had just been given and name peer pressure related to drinking alcohol as one of the five choices.

In general, survey respondents assume that questions on the survey are related to each other, so they will interpret subsequent questions in the context of questions that preceded them. Here is an amusing example.

EXAMPLE 3.4 **Is Happiness Related to Dating?**

Clark and Schober (1992, p. 41) report on a survey that asked the following two questions:

1. How happy are you with life in general?
2. How often do you normally go out on a date? (about ___ times a month)

When the questions were asked in this order, there was almost no relationship between the two answers. But when question 2 was asked first, the answers were highly related. Clark and Schober speculate that in that case, respondents consequently interpreted question 1 to mean; "Now, considering what you just told me about dating, how happy are you with life in general?" ■

Confidentiality versus Anonymity

People sometimes answer questions differently based on the degree to which they believe they are anonymous. Because researchers often need to perform follow-up surveys, it is easier to try to ensure confidentiality than true anonymity. In ensuring confidentiality, the researcher promises not to release identifying information about respondents. In a truly anonymous survey, the researcher does not know the identity of the respondents.

Questions on issues such as sexual behavior and income are particularly difficult because people consider those to be private matters. A variety of techniques have been developed to help ensure confidentiality, but surveys on such issues are hard to conduct accurately.

CASE STUDY 3.1 **No Opinion of Your Own? Let Politics Decide**

SOURCES: Morin (10–16 April 1995), p. 36.
http://www.huffingtonpost.com, 11 April 2013

This is an excellent example of how people will respond to survey questions, even when they do not know about the issues, and how the wording of questions can influence responses. In 1995, the *Washington Post* decided to expand on a 1978 poll taken in Cincinnati, Ohio, in which people were asked whether they "favored or opposed repealing the 1975 Public Affairs Act." There was no such act, but about one-third of the respondents expressed an opinion about it.

In February 1995, the *Washington Post* added this fictitious question to its weekly poll of 1000 randomly selected respondents: "Some people say the 1975 Public Affairs Act should be repealed. Do you agree or disagree that it should be repealed?" Almost half (43%) of the sample expressed an opinion, with 24% agreeing that it should be repealed and 19% disagreeing. The *Post* then tried another trick that produced even more disturbing results. This time, they polled two separate groups of 500 randomly selected adults. The first group was asked: "President Clinton [a Democrat] said that the 1975 Public Affairs Act should be repealed. Do you agree or disagree?" The second group was asked: "The Republicans in Congress said that the 1975 Public Affairs Act should be repealed. Do you agree or disagree?" Respondents were also asked about their party affiliation. Overall, 53% of the respondents expressed an opinion about repealing this fictional act. The results by party affiliation were striking: For the "Clinton" version, 36% of the Democrats but only 16% of the Republicans agreed that the act should be repealed. For the "Republicans in Congress" version, 36% of the Republicans but only 19% of the Democrats agreed that the act should be repealed.

In April 2013, the *Huffington Post* repeated this poll, replacing "Clinton" with "Obama." The results were similar. (Sources: http://www.huffingtonpost.com/2013/04/11/survey-questions-fiction_n_2994363.html and http://big.assets.huffingtonpost.com/toplines_full.pdf) ■

3.3 Open or Closed Questions: Should Choices Be Given?

An **open question** is one in which respondents are allowed to answer in their own words, whereas a **closed question** is one in which they are given a list of alternatives from which to choose their answer. Usually the latter form offers a choice of "other," in which the respondent is allowed to fill in the blank.

Advantages and Disadvantages of Open Questions

As we have seen in Examples 3.1 and 3.3, when people respond to surveys they rarely volunteer answers that aren't among the choices given, even when that option is offered. Therefore, one advantage of open questions is that respondents are free to say whatever they choose, rather than limiting themselves to the choices provided.

The main disadvantage of open questions is that the responses are difficult to summarize. If a survey includes thousands of respondents, it can be a major chore to categorize their responses. Another problem with open questions is that logical responses may not readily come to mind, and the wording of the question might unintentionally exclude answers that would have been appealing had they been included in a list of choices. Schuman and Scott (May 22, 1987) demonstrated this problem with the following example.

EXAMPLE 3.5 **Is the Invention of the Computer Important?**

In the 1980s, Schuman and Scott asked 347 people to "name one or two of the most important national or world event(s) or change(s) during the past 50 years." When asked as an open question, the most common response was World War II (14%), followed by the Vietnam War (10%), the exploration of space (7%), and the assassination of John F. Kennedy (5%). But 54% of respondents didn't give any of those most common answers, and 10% didn't know what to say. Only five people (1.4%) mentioned the invention of the computer, which in retrospect was probably the most important change from that time period. It seems that respondents thought of events and not changes, although the question mentioned both.

The survey was then repeated with a new sample of 354 people who were given a closed form question with five choices. The choices were the four most common ones from the open form, plus "Invention of the computer." When asked as a closed form question, 30% chose the invention of the computer, more than any other response. The common responses from the open form poll also received more support in the closed form question, when they were provided as options. For instance, 23% chose World War II, and 16% chose the exploration of space. ■

Advantages and Disadvantages of Closed Form Questions

One advantage of closed form questions is that they are easier to administer and to analyze than open form questions. Another advantage should be obvious to you if you have taken multiple choice tests—you don't have to come up with the answer on your own; you only need to select from the choices given.

A disadvantage of closed form questions is that they may limit the options because respondents will rarely volunteer a choice that isn't presented, even if they are given the option to choose "Other, please specify." Therefore, it is very important that survey authors think carefully about the choices offered, and it is important that you know what choices were given when you read the results of a survey. Another disadvantage, especially with phone interviews, is that respondents tend to choose the options given later in the list. To compensate for this "recency effect," pollsters often randomize the order of the response options, giving them in different orders to different participants.

Pilot Studies and Pilot Surveys

One compromise that takes into account the advantages and disadvantages of both methods is to conduct an open form **pilot study**, or **pilot survey**, before creating choices for a closed form survey. (No, this doesn't mean that surveys should be conducted on airline pilots!) In a pilot study, a small group of people are asked the questions in open form (the "pilot survey"), and their responses are used to create the choices for the closed form. Often the pilot study will include a focus group discussion to find out what thought process respondents used to arrive at their answers. Other features can be tested in a pilot study as well, such as whether the order of the questions influenced the responses and whether participants understood the questions in the way the survey designers intended.

EXAMPLE 3.6 Questions in Advertising

In her excellent book, *Tainted Truth*, Wall Street Journal reporter Cynthia Crossen explains how advertisers often present results of surveys without giving the full story. As an example, an advertisement for Triumph cigarettes boasted: "TRIUMPH BEATS MERIT—an amazing 60% said Triumph tastes as good as or better than Merit." In truth, three choices were offered to respondents, including "no preference." The results were: 36% preferred Triumph, 40% preferred Merit, and 24% said the brands were equal. So, although the wording of the advertisement is not false, it is also true that 64% said Merit tastes as good as or better than Triumph (Crossen, 1994, pp. 74–75). Which brand do you think wins? ■

CASE STUDY 3.2 **How is the President Supposed to Know What People Think?**

Sources:
http://www.people-press.org/2008/11/13/section-4-early-voting-campaign-outreach-and-the-issues/, released November 13, 2008; accessed May 12, 2013.
http://www.people-press.org/methodology/questionnaire-design/open-and-closed-ended-questions/, accessed May 12, 2013,

If you had just won your first term as president of the United States, you might be curious to know what prompted people to vote for you instead of your opponent. Just get a polling agency to ask them, right? It's not as simple as you might think! In November, 2008, shortly after Barack Obama beat John McCain in the presidential election, the Pew Research organization conducted two polls asking people why they voted as they did. One poll was closed form and the other was open form.

In both the open and closed form of the survey, people were asked "What one issue mattered most to you in deciding how you voted for President?" Results were also compared with an exit poll taken on the day of the election as voters were leaving their polling places. The exit poll was asked in closed form and in person. The other two surveys were conducted by telephone. The results of the open and closed form surveys were strikingly different, as shown in the table below.

"What one issue mattered most to you in deciding how you voted for President?"

Issue:	Open Form	Closed Form	Exit Poll
The economy	35%	58%	63%
The war in Iraq	5%	10%	10%
Health care	4%	8%	9%
Terrorism	6%	8%	9%
Energy policy	0%	6%	7%
Other	43%	8%	0%
Don't know	7%	2%	2%
Total	100%	100%	100%

When presented with the five options shown in the table, over half of respondents chose "The economy." Although "The economy" was still the most frequent response in the open form, only 35% of respondents chose it. Because 43% of respondents in the open choice form chose something not included in the closed form list, all five closed form choices received less support than they did when they were the only options explicitly provided. In case you are curious about what responses were offered by the remaining 43% in the open form, they included things like moral issues, taxes, hope for change, and specific mentions of the candidates by name. ■

Remember that, as the reader, you have an important role in interpreting the results of any survey. You should always be informed as to whether questions were asked in open or closed form, and if the latter, you should be told what the choices were. You should also be told whether "don't know" or "no opinion" was offered as a choice in either case.

3.4 Defining What Is Being Measured

EXAMPLE 3.7 **Teenage Sex**

To understand the results of a survey or an experiment, we need to know exactly what was measured. Consider this example. A letter to advice columnist Ann Landers stated: "According to a report from the University of California at San Francisco . . . sexual activity among adolescents is on the rise. There is no indication that this trend is slowing down or reversing itself." The letter went on to explain that these results were based on a national survey (*Davis* (CA) *Enterprise*, 19 February 1990, p. B-4). On the same day, in the same newspaper, an article entitled "Survey: Americans conservative with sex" reported that "teenage boys are not living up to their reputations. [A study by the Urban Institute in Washington] found that adolescents seem to be having sex less often, with fewer girls and at a later age than teenagers did a decade ago" (p. A-9).

Here we have two apparently conflicting reports on adolescent sexuality, both reported on the same day in the same newspaper. One indicated that teenage sex was on the rise; the other indicated that it was on the decline. Although neither report specified exactly what was measured, the letter to Ann Landers proceeded to note that "national statistics show the average age of first intercourse is 17.2 for females and 16.5 for males." The article stating that adolescent sex was on the decline measured it in terms of frequency. The result was based on interviews with 1880 boys between the ages of 15 and 19, in which "the boys said they had had six sex partners, compared with seven a decade earlier. They reported having had sex an average of three times during the previous month, compared with almost five times in the earlier survey." Thus, it is not enough to note that both surveys were measuring adolescent or teenage sexual behavior. In one case, the author was, at least partially, discussing the age of first intercourse, whereas in the other case the author was discussing the frequency. ■

EXAMPLE 3.8 **Out of Work, Discouraged, but Not Unemployed!**

Ask people whether they know anyone who is unemployed; they will invariably say yes. But most people don't realize that in order to be officially unemployed, and included in the unemployment statistics given by the U.S. government, you must meet very stringent criteria. The Bureau of Labor Statistics uses this definition when computing the official United States unemployment rate (http://www.bls.gov/cps/lfcharacteristics.htm#unemp, accessed May 12, 2013):

> *Persons are classified as unemployed if they do not have a job, have actively looked for work in the prior 4 weeks, and are currently available for work.*

To find the unemployment rate, the number of people who meet this definition is divided by the total number of people "in the labor force," which includes these individuals and people classified as employed. But "discouraged workers" are not included at all. "Discouraged workers" are defined as:

> *Persons not in the labor force who want and are available for a job and who have looked for work sometime in the past 12 months (or since the end of their last job if they held one within the past 12 months), but who are not currently looking because they believe there are no jobs available or there are none for which they would qualify.* (http://www.bls.gov/bls/glossary.htm; accessed May 12, 2013)

If you know someone who fits that definition, you would undoubtedly think of that person as unemployed even though they hadn't looked for work in the past 4 weeks. However, he or she would not be included in the official statistics. You can see that the true number of people who are not working is higher than government statistics indicate. ■

These two examples illustrate that when you read about measurements taken by someone else, you should not automatically assume you are speaking a common language. A precise definition of what is meant by "adolescent sexuality" or "unemployment" should be provided.

Some Concepts Are Hard to Define Precisely

Sometimes it is not the language but the concept itself that is ill-defined. For example, there is still no universal agreement on what should be measured with intelligence, or IQ, tests. The tests were originated at the beginning of the 20th century in order to determine the mental level of school children. The intelligence quotient (IQ) of a child was found by dividing the child's "mental level" by his or her chronological age. The "mental level" was determined by comparing the child's performance on the test with that of a large group of "normal" children, to find the age group the individual's performance matched. Thus, if an 8-year-old child performed as well on the test as a "normal" group of 10-year-old children, he or she would have an IQ of $100 \times (10/8) = 125$.

IQ tests have been expanded and refined since the early days, but they continue to be surrounded by controversy. One reason is that it is very difficult to

define what is meant by intelligence. It is difficult to measure something if you can't even agree on what it is you are trying to measure. If you are interested in knowing more about these tests and the surrounding controversies, you can find numerous books on the subject. A classic book on this topic is by Anastasi and Urbina (1997). It provides a detailed discussion of a large variety of psychological tests, including IQ tests.

EXAMPLE 3.9 Stress in Kids

The studies reported in News Stories 13 and 15 both included "stress" as one of the important measurements used. But they differed in how they measured stress. In Original Source 13, "2003 CASA National Survey of American Attitudes on Substance Abuse VIII: Teens and Parents," teenage respondents were asked:

> *How much stress is there in your life? Think of a scale between 0 and 10, where 0 means you usually have no stress at all and 10 means you usually have a very great deal of stress, which number would you pick to indicate how much stress there is in your life?* (p. 40)

Categorizing responses as low stress (0 to 3), moderate stress (4 to 6), and high stress (7 to 10), the researchers found that low, medium, and high stress were reported by 29%, 45%, and 26% of teens, respectively.

For News Story 15, the children were asked more specific questions to measure stress. According to Additional News Story 15, "To gauge their stress, the children were given a standard questionnaire that included questions like: 'How often have you felt that you couldn't control the important things in your life?'"

There is no way to know which method is more likely to produce an accurate measure of "stress," partly because there is no fixed definition of stress. Stress in one scenario might mean that someone is working hard to finish an exciting project with a tight deadline. In another scenario, it might mean that someone feels helpless and out of control. Those two versions are likely to have very different consequences on someone's health and well-being. What is important is that as a reader, you are informed about how the researchers measured stress in any given study. ■

Measuring Attitudes and Emotions

Similar problems exist with trying to measure attitudes and emotions such as self-esteem and happiness. The most common method for trying to measure such things is to have respondents read statements and determine the extent to which they agree with the statement. For example, a test for measuring happiness might ask respondents to indicate their level of agreement, from "strongly disagree" to "strongly agree," with statements such as "I generally feel optimistic when I get up in the morning." To produce agreement on what is meant by characteristics such as "introversion," psychologists have developed standardized tests that claim to measure those attributes.

3.5 Defining a Common Language

So that we're all speaking a common language for the rest of this book, we need to define some terms. We can perform different manipulations on different types of data, so we need a common understanding of what those types are. Other terms defined in this section are those that are well known in everyday usage but that have a slightly different technical meaning.

Categorical versus Measurement Variables

Thus far in this book, we have seen examples of measuring opinions (such as whether you think abortion should be legal), numerical information (such as weight gain in infants), and attributes that can be transformed into numerical information (such as IQ). To understand what we can do with these measurements, we need definitions to distinguish numerical, quantitative measures from categorical, qualitative ones. Although statisticians make numerous fine distinctions among types of measurements, for our purposes it will be sufficient to distinguish between just two main types: categorical variables and measurement variables. Subcategories of these types will be defined for those who want more detail.

Categorical Variables

Categorical variables are those we can place into a category but that may not have any logical ordering. For example, you could be categorized as male or female. You could also be categorized based on which option you choose as your reason for voting for a particular candidate, as in Case Study 3.2. Notice that we are limited in how we can manipulate this kind of information numerically. For example, we cannot talk about the average reason for voting for a candidate in the same way as we can talk about the average weight gain of infants during the first few days of life.

If the possible categories have a natural ordering, the term **ordinal variable** is sometimes used. For instance, in a public opinion poll respondents may be asked to give an opinion chosen from "strongly agree, agree, neutral, disagree, strongly disagree." Level of education attained may be categorized as "less than high school, high school graduate, college graduate, postgraduate degree." To distinguish them from ordinal variables, categorical variables for which the categories do not have a natural ordering are sometimes called **nominal variables.**

Measurement Variables

Measurement variables, also called **quantitative variables,** are those for which we can record a numerical value and then order respondents according to those values. For example, IQ is a measurement variable because it can be expressed as a single number. An IQ of 130 is higher than an IQ of 100. Age, height, and number of cigarettes smoked per day are other examples of measurement variables. Notice that these can be worked with numerically. Of course, not all numerical summaries will make sense even with measurement variables. For example, if one person in your family smokes 20 cigarettes a day and the remaining three members smoke none, it

is accurate but misleading to say that the average number of cigarettes smoked by your family per day is 5 per person. We will learn about reasonable numerical summaries in Chapter 7.

Interval and Ratio Variables

Occasionally a further distinction is made for measurement variables based on whether ratios make sense. An **interval variable** is a measurement variable in which it makes sense to talk about differences, but not about ratios. Temperature is a good example of an interval variable. If it was 20 degrees last night and it's 40 degrees today, we wouldn't say it is twice as warm today as it was last night. But it would be reasonable to say that it is 20 degrees warmer, and it would mean the same thing as saying that when it's 60 degrees it's 20 degrees warmer than when it's 40 degrees. A **ratio variable** has a meaningful value of zero, and it makes sense to talk about the ratio of one value to another. Pulse rate is a good example. For instance, if your pulse rate is 60 before you exercise and 120 after you exercise, it makes sense to say that your pulse rate doubled during exercise. (And of course having a pulse rate of 0 is *extremely* meaningful!)

Continuous versus Discrete Measurement Variables

Even when we can measure something with a number, we may need to distinguish further whether it can fall on a continuum. A **discrete variable** is one for which you could actually count the possible responses. For example, if we measure the number of automobile accidents on a certain stretch of highway, the answer could be 0, 1, 2, 3, and so on. It could not be 2 1/2 or 3.8. Conversely, a **continuous variable** can be anything within a given interval. Age, for example, falls on a continuum.

Something of a gray area exists between these definitions. For example, if we measure age to the nearest year, it may seem as though it should be called a discrete variable. But the real difference is conceptual. With a discrete variable you can count the possible responses without having to round off. With a continuous variable you can't. In case you are confused by this, note that long ago you probably figured out the difference between the phrases "the number of" and "the amount of." You wouldn't say, "the amount of cigarettes smoked," nor would you say, "the number of water consumed." Discrete variables are analogous to numbers of things, and continuous variables are analogous to amounts. You still need to be careful about wording, however, because we have a tendency to express continuous variables in discrete units. Although you wouldn't say, "the number of water consumed," you might say, "the number of glasses of water consumed." That's why it's the *concept* of number versus amount that you need to think about.

Validity, Reliability, and Bias

The words we define in this section are commonly used in the English language, but they also have specific definitions when applied to measurements. Although these definitions are close to the general usage of the words, to avoid confusion we will spell them out.

Validity

When you talk about something being *valid,* you generally mean that it makes sense to you; it is sound and defensible. The same can be said for a measurement. A **valid measurement** is one that actually measures what it claims to measure. Thus, if you tried to measure happiness with an IQ test, you would not get a valid measure of happiness.

A more realistic example would be trying to determine the selling price of a home. Getting a valid measurement of the actual sales price of a home is tricky because the purchase often involves bargaining on what items are to be left behind by the old owners, what repairs will be made before the house is sold, and so on. These items can change the recorded sales price by thousands of dollars. If we were to define the "selling price" as the price recorded in public records, it may not actually reflect the price the buyer and seller had agreed was the true worth of the home.

To determine whether a measurement is valid, you need to know exactly what was measured. For example, many readers, once they are informed of the definition, do not think the unemployment figures provided by the U.S. government are a valid measure of unemployment, as the term is generally understood. Remember (from Example 3.8) that the figures do not include "discouraged workers." However, the government statistics are a valid measure of the percentage of the "labor force" that is currently "unemployed," according to the precise definitions supplied by the Bureau of Labor Statistics. The problem is that most people do not understand exactly what the government has measured.

Reliability

When we say something or someone is *reliable,* we mean that that thing or person can be depended upon time after time. A reliable car is one that will start every time and get us where we are going without worry. A reliable friend is one who is always there for us, not one who is sometimes too busy to bother with us. Similarly, a **reliable measurement** is one that will give you or anyone else approximately the same result time after time when taken on the same object or individual. In other words, it is *consistent*. For example, a reliable way to define the selling price of a home would be the officially recorded amount. This may not be valid, but it would give us a consistent figure without any ambiguity.

Reliability is a useful concept in psychological and aptitude testing. An IQ test is obviously not much use if it measures the same person's IQ to be 80 one time and 130 the next. Whether we agree that the test is measuring what we really mean by "intelligence" (that is, whether it is really valid), it should at least be reliable enough to give us approximately the same number each time. Commonly used IQ tests are fairly reliable: About two-thirds of the time, taking the test a second time gives a reading within 2 or 3 points of the first test, and, most of the time, it gives a reading within about 5 points.

The most reliable measurements are physical ones taken with a precise measuring instrument. For example, it is much easier to get a reliable measurement of height than of happiness, assuming you have an accurate tape measure. However, you should be cautious of measurements given with greater precision than you think the measuring tool would be capable of providing. The degree of precision probably exceeds the reliability of the measurement. For example, if your friend measures the

width of a swimming pool with a ruler and reports that it is 15.771 feet wide, which is 15′ 9 1/4″, you should be suspicious. It would be very difficult to measure a distance that large reliably with a 12-inch ruler. A second measuring attempt would undoubtedly give a different number.

Bias

A systematic prejudice in one direction is called a *bias*. Similarly, a measurement that is systematically off the mark in the same direction is called a **biased measurement.** If you were trying to weigh yourself with a scale that was not satisfactorily adjusted at the factory and was always a few pounds under, you would get a biased view of your own weight. When we used the term earlier in discussing the wording of questions, we noted that either intentional or unintentional bias could enter into the responses of a poorly worded survey question. Notice that a biased measurement differs from an unreliable measurement because it is consistently off the mark in the same direction.

Connections between Validity, Reliability, and Bias

The differences and connections among the meanings of validity, reliability, and bias can be confusing. For instance, it is not possible for a measurement to be unreliable but still valid in every individual instance. If it's truly measuring what it's supposed to measure every time (i.e., is always valid), it would have to be consistent (i.e. reliable). Let's look at some examples:

- Suppose a woman's weight varies between 140 and 150 pounds, but when asked her weight she always answers (optimistically!) that it's 140 pounds. Then her answer is *reliable*, but it is *not valid* (except on the days when she really does weight 140). Her response is *biased* in the low direction.
- Suppose you take a multiple choice test that truly does measure what you know. Then the test is a *valid* measurement of your knowledge. If you retook a similar test, you should do about equally well, so the test is *reliable*. (Any professor will tell you that designing such tests is a difficult task!)
- A highway patrol officer parks on the side of an open stretch of highway and uses radar to measure the speed of cars as they pass her car. She takes an average of the speeds of 100 cars. The average is a *reliable* measure of the speed of cars passing that point in the sense that she will probably get fairly consistent data, but it is *not a valid measurement* of the average speed of cars for that highway in general. Cars would surely slow down when they saw her parked there, so the measurement is *biased* in the low direction.
- A campus has an academic honesty policy in which students are asked to report all observed cheating incidences to a centralized office. The number of reported incidences per year would probably be relatively consistent and a *reliable* measure of how many incidences students would report, but would *not be a valid measure* of the amount of cheating that occurs. Students are reluctant to turn in other students if they observe them cheating. The measurement would be *biased* in the low direction.
- A thermometer always overestimates the ambient temperature by 5 degrees when it's in the direct sun and always underestimates it by 5 degrees when it's in the shade, but averages out to the correct average daily temperature. At any given

point in time, the thermometer is *not a valid* measure of the ambient temperature, but it does produce a *valid measure* of the daily temperature. As a measuring instrument for daily temperature, the thermometer is *unbiased*. But at any given point in time, it is *biased* either too high (when sunny) or too low (when shady). It is also *reliably* too high or too low at any given time, because it is always off by a consistent amount.

Variability across Measurements

If someone has *variable* moods, we mean that that person has unpredictable swings in mood. When we say the weather is quite variable, we mean it changes without any consistent pattern. Most measurements are prone to some degree of **variability.** By that, we mean that they are likely to differ from one time to the next or from one individual to the next because of unpredictable errors, discrepancies, or natural differences that are not readily explained. If you tried to measure your height as instructed at the beginning of this chapter, you probably found some unexplainable variability from one time to the next. If you tried to measure the length of a table by laying a ruler end to end, you would undoubtedly get a slightly different answer each time.

Unlike the other terms we have defined, which are used to characterize a single measurement, variability is a concept used when we talk about two or more measurements in relation to each other. Sometimes two measurements vary because the measuring device produces unreliable results—for example, when we try to measure a large distance with a small ruler. The amount by which each measurement differs from the true value is called **measurement error.**

Variability can also result from changes across time in the system being measured. For example, even with a very precise measuring device your recorded blood pressure will differ from one moment to the next. Unemployment rates vary from one month to the next because people move in and out of jobs and the workforce. These differences represent **natural variability** across time in the individual or system being measured.

Natural variability also explains why many measurements differ across individuals. Even if we could measure everyone's height precisely, we wouldn't get the same value for everyone because people naturally come in different heights. If we measured unemployment rates in different states of the United States at the same time, they would vary because of natural variability in conditions and individuals across states. If we measure the annual rainfall in one location for each of many years, it will vary because weather conditions naturally differ from one year to the next.

Understanding Natural Variability is the Key to Understanding Statistics

Understanding the concept of natural variability is crucial to understanding modern statistical methods. When we measure the same quantity across several individuals, such as the weight gain of newborn babies, we are bound to get some variability. Although some of this may be due to our measuring instrument, most of it is simply

due to the fact that everyone is different. Variability is plainly inherent in nature. Babies all gain weight at their own pace. If we want to compare the weight gain of a group of babies who have consistently listened to a heartbeat to the weight gain of a group of babies who have not, we first need to know how much variability to expect due to natural causes.

We encountered the idea of natural variability when we discussed comparing resting pulse rates of men and women in Chapter 1. If there were no variability within each sex, it would be easy to detect a difference between males and females. The more variability there is within each group, the more difficult it is to detect a difference between groups. Natural variability can occur when taking repeated measurements on the same individual as well. Even if it could be measured precisely, your pulse rate is not likely to remain constant throughout the day. Some measurements are more likely to exhibit this variability than others. For example, height (if it could be measured precisely) and your opinion on issues like gun control and abortion are likely to remain constant over short time periods.

In summary, variability among measurements can occur for at least the following three reasons:

- Measurements are imprecise, and *measurement error* is a source of variability.
- There is *natural variability across individuals* at any given time.
- There may be *natural variability across time* in a characteristic on the same individual.

The Key to Statistical Discoveries: Comparing Natural Variability to Created Variability

In Part 4, we will learn how to sort out differences due to natural variability from differences due to features we can define, measure, and possibly manipulate, such as variability in blood pressure due to amount of salt consumed, or variability in weight loss due to time spent exercising. In this way, we can study the effects of diet or lifestyle choices on disease, of advertising campaigns on consumer choices, of exercise on weight loss, and so on.

This one basic idea, comparing natural variability to the variability created by different behaviors, interventions, or group memberships, forms the heart of modern statistics. It has allowed Salk to conclude that heartbeats are soothing to infants and the medical community to conclude that aspirin helps prevent heart attacks. We will see numerous other conclusions based on this idea throughout this book.

Thinking About Key Concepts

- Subtle changes in wording, ordering of questions and whether questions are asked as open or closed form can make a big difference in the outcome of a survey.
- People like to please others, so they will give socially acceptable answers in surveys and will even provide opinions on topics they know nothing about, pretending that they do know.