



chapter

2



Smithsonian American Art Museum, Washington, DC/Art Resource, NY

# The Research Process: *How We Find Things Out*

## chapter outline

### **The Scientific Method: Designed to Be Valid**

*Step 1: Specifying a Problem*

*Step 2: Observing Events*

*Step 3: Forming a Hypothesis*

*Step 4: Testing the Hypothesis*

*Step 5: Formulating a Theory*

*Step 6: Testing a Theory*

### **The Psychologist's Toolbox: Techniques of Scientific Research**

*Experimental Research*

*Correlational Research*

*Descriptive Research*

*Be a Critical Consumer of  
Psychology*

### **Statistics: Measuring Reality**

*Descriptive Statistics*

*Inferential Statistics*

*Lying With Statistics*

### **How to Think About Research Studies**

*Reading Research Reports:  
The QALMRI Method*

*Writing Your Own Research Papers*

**I**n 1984, President Ronald Reagan gave his official blessing to a daring project, building a permanent inhabited space station. A key part of President Reagan's vision was that the space station should be built and staffed by people from many different countries. His vision came to pass, and the International Space Station (ISS) is now being constructed some 250 miles above earth. The ISS is a mammoth project, comparable in size and scope to the pyramids of ancient Egypt. The station will span a distance greater than a football field and will weigh over 1 million pounds. Its solar panels will spread over almost an acre, and it will cost at least \$90 billion.

The ISS is not just a technological marvel, it is also a testament to our very human ability to cooperate and interact effectively. The project is particularly impressive because it is being built by 16 countries—including former enemies, notably the United States and Russia.

In the first phase of the program (which began in 1995), American and Russian astronauts lived and worked together on the Russian *Mir* space station. These experiences taught astronauts and earth-bound scientists about living and working in space and, equally important, they also built cooperation and trust between the astronauts themselves and the respective organizations back on earth.

The ISS will be used not only to observe the earth, but also to experiment with new ways to manufacture materials, to make drugs, and to study diseases ranging from osteoporosis to cancer. But



The ISS is not just a technological marvel, it is also a testament to our very human ability to cooperate and interact effectively.

study diseases ranging from osteoporosis to cancer. But more than that, the science of psychology will play a key role in discovering the best ways for people to live and function in space. The ISS will allow us to understand new facets of human mental processes and behavior, both when people are alone and when they are part of a group. How can we learn such things from the ISS, or from events in any other context? Through science. In this chapter, we see how the science of psychology can be used to learn about the brain, the person, and the group.

## The Scientific Method: Designed to Be Valid

On November 2, 2000, the first astronauts moved into the ISS. They moved into just one small part of this work-in-progress, which provided the bare minimum in basic facilities. Among these basic essentials were a treadmill and an exercise bike. These pieces of equipment are not luxury items, but absolute necessities: Astronauts must exercise to retain their bone density and muscles. The trouble is (as too many of we earth-bound couch potatoes know only too well) that the benefits of exercising occur later, whereas the benefits of coping with the press of immediate events occurs right now—and thus it's sometimes difficult to be motivated to exercise. A substitute has been proposed: Simply vibrating the bones appears to prevent calcium loss. But could this procedure also disrupt the astronauts' concentration? Perhaps make it difficult for them to sleep?

How could we find out whether the new vibration treatment would have unintended consequences, for example by affecting the quality of sleep (either as the procedure was conducted, or afterward)? Psychology is a science because it relies on a specific type of method of inquiry, and this method, in principle, allows us to discover characteristics that predict human behavior.

The **scientific method** is a way to gather facts that will lead to the formulation and validation of a theory. It involves *specifying a problem; systematically observing events; forming a hypothesis of the relation between variables; collecting new observations to test the hypothesis; using such evidence to formulate and support a theory; and finally, testing the theory*. Let's take a closer look at the scientific method, one step at a time.

● **Scientific method:** The scientific method involves specifying a problem, systematically observing events, forming a hypothesis of the relation between variables, collecting new observations to test the hypothesis, using such evidence to formulate and support a theory, and finally testing the theory.

### Step 1: Specifying a Problem

What do we mean by “specifying a problem”? Science tries to answer questions, any one of which may be rephrased as a “problem.” Despite the way the word

is often used in ordinary conversation, a problem is not necessarily bad: It is simply a question you want to answer, or a puzzle you want to solve. A scientist might notice what seems to be a consistent pattern, for example, and wonder if it reflects a connection or a coincidence. A scientist might notice that many cultures have an afternoon siesta or “tea time” and wonder whether humans have a biological rhythm that makes us sluggish at that time of day—and might then ask how living in weightlessness affects such biological rhythms. Speaking metaphorically, a scientist might notice a nail sticking above the boards and ask why it is different from other nails. Selecting astronauts is an example. We ask *what* personal characteristics will lead someone to be effective in this job, *why* those characteristics will lead them to behave in certain ways in specific circumstances, and *how* those characteristics were acquired and can be further developed.

## Step 2: Observing Events

Consider the idea of “systematically observing events.” Scientists are not content to rely on impressions or interpretations. They want to know the facts, as free from any particular notions of their significance as possible. Facts are established by collecting **data**, which are numerical measurements or careful observations of a phenomenon. Properly collected data can be **replicated**, that is, collected again by the original investigator or someone else. Scientists often prefer quantitative data (numerical measurements), such as how quickly a person can respond to an unexpected event or how accurately a person can establish up versus down in weightlessness. In addition to collecting numerical data, scientists rely on systematic observations, which simply document that a certain event occurs. But unless the data include numbers, it is often difficult to sort the observation from the interpretation. The data are just the facts, ma’am, nothing but the facts—when data are collected the interpretation must be set aside, saved for later.

What do we mean by “events”? An event in the scientific sense is the occurrence of a particular phenomenon. Scientists study two kinds of events: those that are themselves directly observable (such as how many times in an hour a mother strokes her infant) and those that, like thoughts, motivations, or emotions, can only be inferred. For example, when people smile without really meaning it (as many people do when posing for photographs), the muscles used are not the same ones that produce a sincere smile (Ekman, 1985). It is possible to observe directly which set of muscles is in use, and the recorded data would distinguish between the two kinds of contractions. But the researcher’s interest goes beyond the directly observable muscle contractions to the link with inner (and invisible) thoughts and feelings. By studying what’s observable (the muscles), researchers can learn about the unobservable (the mental state of the smiler).

## Step 3: Forming a Hypothesis

What about “forming a hypothesis of the relation between variables”? First, by the term **variable**, researchers mean an aspect of a situation that is liable to change (or, in other words, that can vary); more precisely, a variable is a characteristic of a substance, quantity, or entity that is measurable. A **hypothesis** is a tentative idea that might explain a set of observations. For example, the ISS does not have normal patterns of “day” and “night,” and thus astronauts do not sleep when it gets dark outside. Rather, they may get wrapped up in what they are doing and forget how much time has passed since they last slept. Could losing sleep interfere with learning? For

- **Data:** Objective observations.
- **Replication:** Collecting the same observations or measurements and finding the same results as were found previously.
- **Variable:** An aspect of a situation that can vary, or change—specifically, a characteristic of a substance, quantity, or entity that is measurable.
- **Hypothesis:** A tentative idea that might explain a set of observations.

optimal performance on the ISS, it's important to know whether lack of sleep impairs learning and later memory; memory lapses that would lead to harmless mishaps on earth could lead to disaster in space. This hypothesis comes down to the assertion that there's a connection between two variables—not sleeping and learning. However, before you urge the ISS astronauts to force themselves to sleep 8 hours in 24, you ought to test the hypothesis to find out whether it's correct.

### Step 4: Testing the Hypothesis

Thus, you must go about “collecting new observations to test the hypothesis.” The first thing you need to do is create operational definitions of the key concepts, which makes them concrete enough to test. An **operational definition** specifies a variable by indicating how it is measured or manipulated. In this example, “not sleeping” might be defined as having stayed continuously awake for 24 hours, and learning might be defined as retaining memory for material that was studied earlier in the day. Are there rigorous studies that bear on memory following a sleepless night? Yes, and researchers have found that staying up all night does disrupt memory for information learned that day (Graves et al., 2001; Stickgold et al., 2000). In fact, memory for verbal material is impaired even if people sleep, when their sleep patterns are disrupted because they've had to wake up intermittently (Ficca et al., 2000). A typical study has two groups: In one, participants learn some information, such as a list of words. These people then sleep normally, and their memory is tested the next day. Participants in a second group learn the same material (and learn it as well) as the participants in the first group, but stay up all night and are tested the next day. The hypothesis is that the participants will have better memory if they were allowed to sleep.

### Step 5: Formulating a Theory

Now consider “using such evidence to formulate and support a theory.” A **theory** consists of an interlocking set of concepts or principles that explains a set of observations. Unlike a hypothesis, a theory is not a tentative idea and doesn't focus on possible relationships among variables. Instead, theories are rooted in an established web of facts and concepts and focus on the *reasons* for established relationships among variables. In our example, the notion that people will fail to store information in memory if they don't sleep is a hypothesis, not a theory. A theory might explain that sleep is necessary for learning because: (1) Specific brain areas are activated when we learn during the day; and (2) Those areas must continue to operate for a specific period of time while we sleep in order to store the information acquired during the day. Hypotheses and theories both produce **predictions**, expectations about specific events that should occur in particular circumstances if the hypothesis or theory is correct.

### Step 6: Testing a Theory

Finally, what do we mean by “testing the theory”? The history of science is littered with theories that turned out to be wrong. Researchers evaluate a theory by testing its predictions. For example, in one study researchers scanned the brains of people as they learned sequences of responses, and then scanned their brains again while they slept that night. Brain areas that were active during learning continued to be active during sleep. Moreover, these areas were more active during sleep for the people who had learned the task that day than for others who had not learned it (Maquet et al., 2000).

● **Operational definition:** A definition of a variable that specifies how it is measured or manipulated.

● **Theory:** An interlocking set of concepts or principles that explain a set of observations.

● **Prediction:** An expectation about specific events that should occur in particular circumstances if the theory or hypothesis is correct.

Each time a theory makes a correct prediction, the theory is supported, and each time it fails to make a correct prediction, the theory is weakened. If enough of its predictions are unsupported, the theory must be rejected and the data explained in some other way. A good theory is *falsifiable*; that is, it makes predictions it cannot “squirm out of.” A falsifiable theory can be rejected if the predictions are not confirmed.



## Looking at Levels

### Sleep Disturbances and Cognitive Function

Sleep apnea (“apnea” in Greek means “without breath”) is a disorder in which people stop breathing while they sleep. When the blood oxygen level falls low enough, brain circuits are activated to wake the person and to restart breathing. About 12 million Americans have this problem (American Sleep Apnea Association, 2002, <http://www.sleepapnea.org/index2.html>); some victims suffer hundreds of episodes a night. In most cases, the problem is caused by an obstruction in the throat, which also leads the person to snore while sleeping. Researchers hypothesized that because sleep apnea disrupts sleep, it should also impair memory and other cognitive abilities. To test this hypothesis, they studied the effects that surgically removing the obstruction had on learning, memory, and decision-making in 53 people who had this disorder. They administered tests of cognitive abilities before surgery and 6 months after surgery. Happily, surgery did lead to better learning, memory, and decision-making abilities. Moreover, the researchers found that the more effective the treatment was in eliminating the apnea, the larger the improvement in these abilities (Dahloef et al., 2002).

Consider this finding from the perspective of levels of analysis. First, brain processes monitor oxygen levels in the blood and wake the person. If sleep is disrupted, so is the ability to retain information acquired during the day. Second, at the level of the person, people with apnea typically feel drowsy during the day. They find it difficult to sustain interest and motivation. Third, at the level of the group, they under-perform at work, and are not as helpful to co-workers as they could be. In addition, these people typically snore as they sleep, which can disrupt the sleep patterns of their partners—impairing their partners’ cognitive abilities the next day. (Even though sleep apnea is not contagious, its effects may be!) The events at the three levels of analysis interact. Indeed, relationships can be strained if one partner snores, and that strain can itself contribute to sleep problems, which in turn could make the cognitive deficits even worse. And consider the role of the surgeon: The social interactions between patient and surgeon, which culminate in the patient’s deciding to have surgery, can have dramatically positive effects—vastly improving the quality of life of the sufferer.

#### TEST YOURSELF!

1. What are the steps of the scientific method?

## The Psychologist’s Toolbox: Techniques of Scientific Research

Life on the ISS is not like life on earth. The beginning of the day isn’t signaled by the dawn’s early light, and your morning shower is more likely to be a sponge bath. Drinking from a glass is a challenge; a wrong nudge, and a glob of liquid floats up into your face—or, worse yet, escapes and drifts toward sensitive equipment mounted on a nearby wall or ceiling. And forget about sleeping in a bed. Instead, you crawl into a special sleeping bag that is anchored to a wall, and tie your arms

● **Independent variable:** The aspect of the situation that is intentionally varied while another aspect is measured.

● **Dependent variable:** The aspect of the situation that is measured as an independent variable is changed. The value of the dependent variable *depends* on the independent variable.

down before nodding off. But perhaps the most striking differences from life as we usually experience it are social. The ISS will only house 7 people. How will they get along as time goes on? How should living quarters be designed to ensure enough privacy, to “give them some space”—but at the same time promote social support? How should daily wake/sleep schedules be synchronized for the different crew members? And how many of the crew should be men, how many women? Would it be best if most of the crew were married couples?

The questions go on and on. How can we answer them? Although all sound psychological investigations rely on the scientific method, the different areas of psychology often pose and answer questions differently. Psychologists use a variety of research tools, each with its own advantages and disadvantages.

## Experimental Research

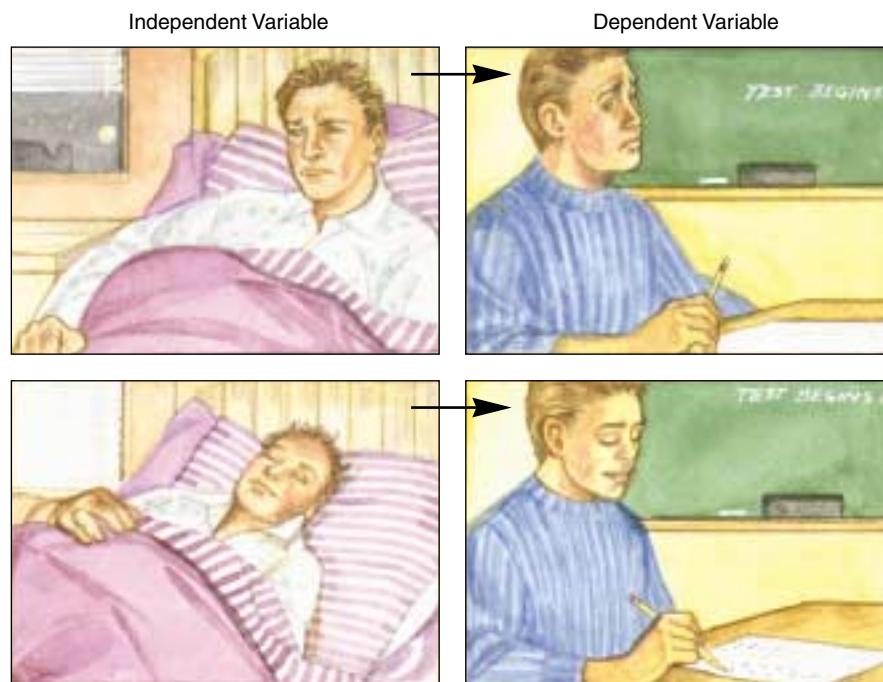
Much psychological research relies on conducting experiments, controlled situations in which variables are manipulated.

### Independent and Dependent Variables

The variables in a situation—for example, “amount of sleep” and “memory performance”—are the aspects of a situation that can vary. The experimenter deliberately alters one aspect of a situation, which is called the **independent variable**, and measures another, called the **dependent variable**. In other words, the value of the dependent variable depends on the value of the independent variable. In our sleeplessness and memory example in the previous section, whether or not participants slept was the independent variable (it was deliberately varied), and memory performance was the dependent variable (it was measured; see Figure 2.1). By examining

**FIGURE 2.1** Relationship Between Independent and Dependent Variables

The independent variable is what is manipulated—whether or not participants slept. In this example, the dependent variable, what is measured, is memory performance the next day.



the link between independent and dependent variables, a researcher hopes to discover exactly which factor is causing an **effect**, which is the difference in the dependent variable that results from a change in the independent variable. In our sleeplessness and memory example, the effect is the degree to which memory is better following sleep than it is following a sleepless night.

Once researchers have found a relation between two variables, they need to test that relation to rule out other possible explanations for it; only by eliminating other accounts can we know whether a hypothesized relation is correct. Say we had tested only one group, the one that was kept awake. The fact that these students failed to recall many words would not necessarily show that sleep is critical for learning. Why? Perhaps the test was too difficult and, even in the best of circumstances, the participants wouldn't have been able to remember much. Or perhaps the testing situation created a lot of anxiety, and that's what interfered with learning. A **confound**, or *confounding variable*, is another possible aspect of the situation (such as the anxiety that accompanies a test) that has become entangled with the aspects that you have chosen to vary. Confounds thus lead to results that are ambiguous, that do not have a clear-cut interpretation (see Figure 2.2).

### Experimental and Control Conditions

One way to disentangle confounds is to use a control group. A **control group** is treated identically to the experimental group except with regard to the one variable you want to study; a good control group holds constant—or controls—all of the variables in the experimental group except the one of interest. In experiments on the role of sleep in learning, the experimental group doesn't sleep; the control group does. If the kinds of people assigned to the two groups differ markedly, say, in age, gender, or learning ability (or any combination of the three), those factors could be confounds that would mask a clear reading of the experiment's results; any difference in the groups' performance could have been caused by any of those elements. For instance, if the sleepless group happened to include more people

- **Effect:** The difference in the dependent variable that is due to the changes in the independent variable.

- **Confound** (or *confounding variable*): An independent variable that varies along with the ones of interest, and could be the actual basis for what you are measuring.

- **Control group:** A group that is treated exactly the same way as the experimental group, except for the one aspect of the situation being studied—the independent variable. The control group holds constant—"controls"—all of the variables in the experimental group except the one of interest.

**FIGURE 2.2** Confounding Variable in Everyday Life



A particularly dramatic example of a confounding variable in everyday life was present during the Victorian age. At this time in history, women often were considered frail and delicate creatures, at least partly because they seemed prone to fainting spells. Did they faint because of their “inner natures” or for some other reason? Consider the fact that many of these women wore extremely tight corsets to give them tiny waists. In fact, the corsets were so tight that women could only take shallow breaths—if they took a deep breath, they ran the risk of being stabbed by the whalebone “stays” in the corset. These stays were thin and very sharp, and not only could they cause a bloody wound, but they could also puncture a lung! One consequence of continued shallow breathing is dizziness—hence the fainting spells common among stylish Victorian women. Can you think of an experiment to show conclusively that the corsets were to blame? How do you think views of women would have been different if men had worn tight corsets too?

who were elderly than the control group, the researchers should not conclude that sleeplessness led the experimental group to forget—perhaps these people simply learned less effectively in general. In a properly conducted experiment, therefore, the researchers rely on **random assignment**: participants are assigned randomly, that is, by chance, to the experimental and the control groups, so that no confounds can sneak into the composition of the groups.

Similarly, you can use a **control condition**, either for a group of people or a single person. Instead of testing a separate control group, you test the same group another time, keeping everything the same as in the experimental condition except for the single independent variable of interest. For example, you could test the same people twice, once when they were allowed to sleep normally after studying the words and once when they were kept awake. (Indeed, you can test them four times, twice while on earth and then twice while on the ISS; this experiment would allow you to discover whether sleep has the same effects on memory in weightlessness as it does normally.) To avoid confounding the order of testing with the condition (experimental versus control), you would test half the participants in the control condition before testing them in the experimental condition and would test the other half of the participants in the experimental condition before testing them in the control condition.

### Quasi-Experimental Design

One element of a true experiment is that the participants are assigned randomly to the different groups. But in the real world, it is not always possible or desirable to achieve randomness, and so sometimes research designs must be *quasi-experimental* (quasi means “as if” in Latin). For instance, let’s say that you want to discover whether the effects of sleep on learning are different for people of different ages, so you decide to test four groups of people: teenagers, college students, middle-aged people, and the elderly. Obviously you cannot assign people to the different age groups randomly. You should control for as many variables—such as health and education level—as you can in order to make the groups as similar as possible. Similarly, if you want to track changes over time (for example, in astronauts’ memory abilities after they return from a stint in the ISS), it is not possible to assign people randomly to the groups as time goes by because you are taking measurements only from people you have measured before. In these examples, participants are not assigned randomly to groups, and such quasi-experiments rely on comparing multiple groups or multiple sets of measurements, attempting to eliminate potential confounds as much as possible. Unfortunately, because the groups can never be perfectly equated on all characteristics, you can never be certain exactly what differences among groups are responsible for the observed results. The conclusions you draw from quasi-experiments cannot be as strong as those from genuine experiments.

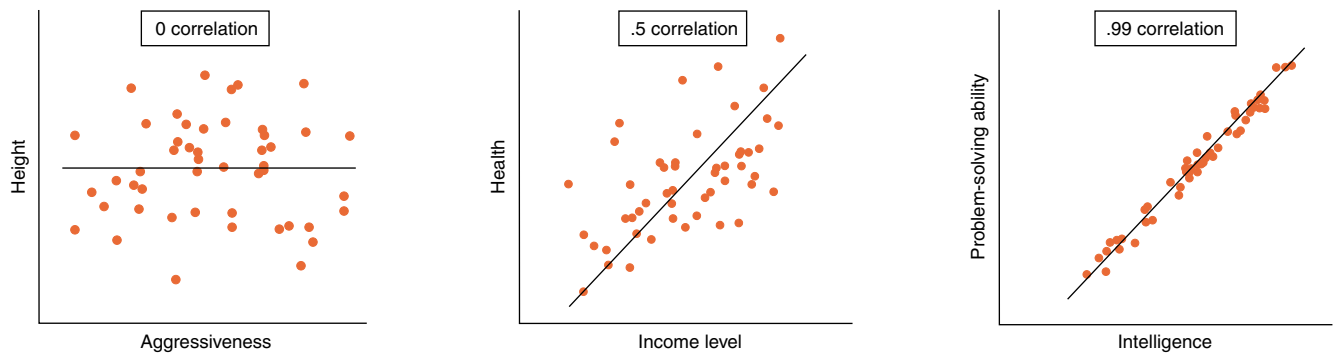
- **Random assignment:** Participants are assigned randomly, that is, by chance, to the experimental and the control groups, so that no biases can sneak into the composition of the groups.

- **Control condition:** Like a control group, but administered to the same participants who receive the experimental condition.

## Correlational Research

Sometimes it is not possible to do an experiment or even a quasi-experiment, particularly if you are interested in studying large groups or if it is difficult (or unethical) to manipulate the variables. Let’s say you want to answer this question: Do people who need minimal sleep make better astronauts? Not only is there a problem in randomly assigning participants to the long-sleep and minimal-sleep groups, but you also can’t simply declare that someone is an astronaut. For a question like this, not even a quasi-experiment can be performed.

## FIGURE 2.3 Strength of Correlation



Tall people are as likely to be aggressive or gentle as short people, which is indicated by the 0 correlation between the two variables.

The more tightly the dots cluster around the line, the higher the correlation; the higher the correlation, the more accurately the value on one dimension predicts the value on the other. Here we see that knowing someone's income level predicts something about how healthy they are likely to be, but not very much.

We would expect smarter people to be better at solving problems, as reflected in this very high correlation. Although these are fictional data, why might you make these predictions?

In such situations, researchers use another method to study the relations among variables, a method that relies on the idea of correlation. **Correlation** is a relationship in which changes in the measurements of one variable are accompanied by changes in the measurements of another variable. A *correlation coefficient* (often simply called a *correlation*) is an index of how closely related two measured variables are. Figure 2.3 illustrates three predicted correlations between variables. A positive relationship (in which increases in one variable are accompanied by increases in another) is indicated by a correlation value that falls between zero and 1.0; a negative relationship (in which increases in one variable are accompanied by decreases in another) is indicated by a correlation that is between zero and  $-1.0$ . If we had plotted “sickness” instead of “health” in the middle panel, the line would have gone down—and the correlation we plotted would have been  $-0.5$  instead of  $+0.5$ . A zero correlation indicates no relationship between the two variables; they do not vary together. The closer the correlation is to 1.0 or  $-1.0$ , the stronger the relationship; visually, the more tightly the numbers cluster around the line, the higher the correlation.

Correlational research involves measuring at least two things about each of a set of individuals or groups (or measuring the same individuals or groups at a number of different times), and looking at the way one set of measurements goes up or down in tandem with another set of measurements; correlations always compare one pair of measurements at a time. The main advantage of correlational research is that it allows researchers to compare variables that cannot be manipulated directly. The main disadvantage is that correlations indicate only that two variables tend to vary together, not that one *causes* the other. For example, evidence suggests a small correlation between poor eyesight and intelligence (Belkin & Rosner,

● **Correlation:** An index of how closely interrelated two sets of measured variables are, which ranges from  $-1$  to  $+1$ . The higher the correlation (in either direction), the better you can predict the value of one type of measurement when given the value of the other.



Researchers have found that the lower the level of a chemical called monoamine oxidase (MAO) in the blood, the more the person will tend to seek out thrilling activities (such as sky diving and bungee jumping; Zuckerman, 1995). Thus there is a negative correlation between the two measures: as MAO levels go down, thrill seeking goes up. But we don't know whether MAO level causes the behavior or vice versa—or whether some other chemical, personality trait, or social factor causes the levels of both MAO and thrill seeking to vary together.

Some scientists observe animals in the wilds of Africa; others observe sea life in the depths of the ocean; and others observe humans in their natural habitats.

1987; Miller, 1992; Williams et al., 1988), but poor eyesight doesn't cause someone to be smarter! Similarly, researchers have found that weightlessness disrupts spatial orientation (such as awareness of the position of your body) and that these effects may be related to space motion sickness (Young et al., 1993). Does motion sickness disrupt spatial orientation, or does impaired spatial orientation produce motion sickness? Or, does some other variable—such as abnormal head movements—produce both effects? Just given the correlation between problems in spatial orientation and the occurrence of motion sickness, you can't say. Remember: *Correlation does not imply causation*. In contrast, in

an experiment, you can manipulate the independent variable and hold everything else constant, and thereby show that changes in the independent variable cause changes in the dependent variable.

## Descriptive Research

Although the scientific method is always described in terms of testing hypotheses, this isn't quite the whole story. Not all research is sparked by specific hypotheses. Some research is devoted simply to describing “things as they are.” Theorizing without facts is a little like cooking without ingredients.

### Naturalistic Observation

For the scientist, “facts” are not intuitions, impressions, or anecdotes. Essential to the scientific method is careful, systematic, and unbiased observation that can be repeated by others, and some researchers specialize in collecting such data from real-world settings. For example, some researchers observed caregivers interacting with young children, and noted that the caregivers changed their language and speech patterns, using short sentences and speaking in a high pitch. This speech modification, originally dubbed *motherese*, is now often called *child-directed speech* (Morgan & Demuth, 1996; Snow, 1991, 1999).



Although naturalistic observation is an essential part of science, it is only the first step. You cannot control for confounded variables, and you cannot change the variables to see what the critical factor in a particular mental process or behavior might be. The discovery of motherese does not tell us whether caregivers use it in order to help children understand

them, or to entertain them, or simply to imitate other caregivers they have heard. In science, observing an event is only the first step.

### Case Studies

Sometimes nature or human affairs lead to unique situations, which change an independent variable in a novel way. A **case study** focuses on a single instance of a situation, examining it in detail. For example, a researcher might study a single astronaut, looking closely at her life and circumstances in an effort to formulate hypotheses about the psychological underpinnings that lead someone to succeed in this profession. Many neuropsychologists study individual brain-damaged patients in depth to discover which abilities are “knocked out” following certain types of damage (in Chapter 3 you will read about a young soldier who, after suffering brain injury, had bizarre visual impairments). A psychologist who studies abnormal behavior might study a reported case of multiple personalities to discover whether there’s anything to the idea (books such as *Sybil* and *The Three Faces of Eve* describe such cases in great detail), a cognitive psychologist may investigate how an unusually gifted memory expert is able to retain huge amounts of information almost perfectly, and a personality psychologist might study in detail how astronauts remain motivated through years and years of hard work with no guarantee that they will ever fly into space.

Unlike naturalistic observation, case studies are not necessarily limited simply to describing values of variables and relations among them. Rather, in some situations you can actually perform experiments with only a single case (such as testing the ability of a memory expert to recall lists of familiar versus unfamiliar things, or to remember words versus pictures), and such studies can help us understand the particular situation in detail. Nevertheless, we must always be cautious about generalizing from a single case; that is, we must be careful in assuming that the findings in the case study extend to all other similar cases. Any particular person may be unusual for many reasons and so may not be at all representative of people in general.

### Surveys

A **survey** is a set of questions put to the participants about their beliefs, attitudes, preferences, or activities. Surveys are a relatively inexpensive way to collect a lot of data fairly quickly, and they are popular among psychologists who study personality and social interactions. Surveys provide data that can be used to formulate or test a hypothesis. However, the value of surveys is limited by what people are capable of reporting accurately. You could ask people how they feel about the government’s using tax dollars to build the ISS, but would not use a survey to ask people how their brains work, or to report subtle behaviors, such as body language,



Brain damage following an accident can cause someone to fail to name fruits and vegetables while still able to name other objects (Hart et al., 1985). A case study would examine such a person in detail, documenting precisely what sorts of things could and could not be named.

- **Case study:** A scientific study that focuses on a single instance of a situation, examining it in detail.
- **Survey:** A set of questions, typically about beliefs, attitudes, preferences, or activities.

that they may engage in unconsciously. Moreover, even if they are capable of answering, people may not always respond honestly; as we note in Chapter 11, this is especially a problem when the survey touches on sensitive personal issues, such as sex. And even if people do respond honestly, what they say does not always reflect what they do. We cite a classic example in Chapter 16, reported in 1934 by La Piere, in which restaurant managers were asked whether they would serve Chinese people; although most said they would not, when Chinese people actually came to their restaurants, virtually all served them without question. Finally, not everyone who is asked to respond does, in fact, fill in the survey. Because a particular factor (such as income or age) may incline some people, but not others, to respond, it is difficult to know whether you are justified in generalizing from the respondents to the rest of the group of interest.

Survey questions have to be carefully worded so that they don't lead the respondent to answer in a certain way and yet still get at the data of interest. In the survey mentioned in Chapter 1, in which the respondents had to decide which was worse, a mate's having sex with or becoming emotionally involved with someone else, people who disliked both possibilities equally had no way to say so, and the forced choice would not reflect their true position (which may be a serious problem with this study; DeSteno & Solvey, 1996; Harris, 2002). Similarly, the nature of the response scale (for example, the range of values presented) affects what people say, as does the order in which questions are asked (Schwarz, 1999). Table 2.1 summarizes the basic research methods used in psychology.

**TABLE 2.1** Summary of Research Methods

<b>Experimental design</b>	Participants are assigned randomly to groups, and the effects of manipulating one or more independent variables on a dependent variable are studied.
<b>Quasi-experimental design</b>	Similar to experiments but participants are not assigned to groups randomly.
<b>Correlational research</b>	Relations among different variables are documented, but causation cannot be inferred.
<b>Naturalistic observation</b>	Observed events are carefully documented.
<b>Case study</b>	A single instance of a situation is analyzed in depth.
<b>Survey</b>	Investigation requires participants to answer specific questions.

## Be a Critical Consumer of Psychology

No technique is always used perfectly, so you must be a critical consumer of all science, including the science of psychology. Metaphorically speaking, there are no good psychologists on salt-free diets—we take everything with at least a grain of salt! But this doesn't mean that you should be cynical, doubting everything you hear or read. Rather, whenever you read a report of a psychological finding in a

newspaper, a journal article, or a book (including this one), look for aspects of the study that could lead to alternative explanations. You already know about the possibility of confounds; here are a few other issues that can cloud the interpretation of studies.

**Reliability: Count on It!**

Not all data are created equal; some are better than others. One way to evaluate data is in terms of reliability. **Reliability** means consistency. A reliable car is one you can count on to behave consistently, starting even on cold mornings and not dropping random parts on the highway. A reliable study is one that can be replicated, that is, repeated with the same results. When you read about the results of a study, find out if they have been replicated; if so, then you can have greater confidence that the results are reliable.

**Validity: What Does It Really Mean?**

Something is said to be valid if it is what it claims to be; a valid driver’s license, for example, is one that was, in fact, issued by the state and has not expired (and thus does confer the right to drive). In science, **validity** means that a method provides a true measure of what it is supposed to measure. A study may be reliable but not valid, or vice versa. Table 2.2 lists four of the major types of validity (Carmines & Zeller, 1979).

● **Reliability.** Data are reliable if the same results are obtained when the measurements are repeated.

● **Validity:** A measure is valid if it does in fact measure what it is supposed to measure.

**TABLE 2.2** Four Major Types of Validity

Type	Description	Example
Face validity	Design and procedure appear to assess the variables of interest.	Sample essay as part of an entrance exam for journalism school.
Content validity	Measures assess all aspects of phenomenon of interest.	Test of knowledge of research methods that covers all methods.
Criterion validity	A measure or procedure is comparable to a different, valid measure or procedure.	A paper-and-pencil test of leadership ability correlates highly with poll results of leadership of actual leaders.
Construct validity	Measures assess variables specified by a theory.	A theory defines “fatigue” in terms of lack of alertness, and the measure assesses this lack.

To understand the concept of validity, let’s see what it’s like to be a participant in a study. So, before reading further, try this exercise. Table 2.3 (p. 40) contains a list of words. Decide whether the first word names a living object or a nonliving one (circle the word “living” at the right if it is living; otherwise move to the next word); then decide whether the second word begins with the letter t (circle the words “begins with t” if it does; otherwise move to the next word); then decide whether the third word names a living or a nonliving object, whether the fourth word begins with the letter t, and so on, alternating judgments as you go down the list. Please do this now.



### TABLE 2.3 What's in a Word?

Circle the word on the right if the word on the left has the named property; otherwise, move on to the next word. After you finish the list, read on.

salmon	living	trout	living
tortoise	begins with <i>t</i>	donkey	begins with <i>t</i>
airplane	living	teapot	living
toad	begins with <i>t</i>	house	begins with <i>t</i>
guitar	living	table	living
goat	begins with <i>t</i>	terrain	begins with <i>t</i>
truck	living	tiger	living
automobile	begins with <i>t</i>	rosebush	begins with <i>t</i>
snake	living	bacteria	living
tent	begins with <i>t</i>	carpet	begins with <i>t</i>
toast	living	staple	living
television	begins with <i>t</i>	tricycle	begins with <i>t</i>
wagon	living	lawn	living
tarantula	begins with <i>t</i>	ocean	begins with <i>t</i>
toadstool	living	tuna	living
elephant	begins with <i>t</i>	terrier	begins with <i>t</i>

When you have finished marking the list, take out a piece of paper and (without looking!) write down as many of the words as you can. How many words from the list were you able to remember?

The standard result from this kind of study is that people will remember more words after making a living/nonliving judgment than after making a *t*/non-*t* judgment (for example, see Craik & Tulving, 1975). This result is usually interpreted to mean that the more we think about (or “process”) the material, as we must in order to make the living/nonliving decision, the better we remember it; for the *t* words, we only need to look at the first letter, not think about the named object at all. In fact, if we are forced to think about something in detail but don’t consciously try to learn it, we end up remembering it about as well as if we did try to learn it (we discuss this curiosity more in Chapter 7).

Does this demonstration of differences in memory following differences in judgment really bear out this interpretation? What if you remembered the words you judged as living/nonliving better because you had to read the whole word to make the required judgment, but you only looked at the first letter of the other words to decide whether they began with *t*? If this were the case, your better memory of words in the living/nonliving category would have nothing to do with “think-



A study of mind-reading can be highly reliable (repeatable), but not valid if the mind-reader has figured out a way to cheat; and a study that uses facial expression to determine what someone is feeling can be valid but not reliable (because a given expression may be difficult to discriminate from other expressions).

ing about it more.” Therefore, the experiment would not be valid—it would not be measuring what the investigator designed it to measure.

When you read a result, always try to think of as many interpretations for it as you can; you may be surprised at how easy this can be. And, if you can think of an alternative interpretation, see whether you can think of a control group or condition that would allow you to tell who was right, you or the authors of the study.

### **Bias: Playing With Loaded Dice**

Sometimes beliefs, expectations, or habits alter how participants in a study respond or affect how a researcher sets up or conducts a study, thereby influencing its outcome. This leaning toward a particular result, whether conscious or unconscious, is called **bias**, and it can take many forms. One form of bias is **response bias**, in which people have a tendency to respond in a particular way regardless of their actual knowledge or beliefs. For example, many people tend to say yes more than no, particularly in some Asian cultures (such as that of Japan). This sort of bias toward responding in “acceptable” ways is a devilish problem for survey research. For example, when asked whether you support research on the ISS that could produce a cure for cancer, you would be hard-pressed to say “no.” Another form of bias is **sampling bias**, which occurs when the participants or items are not chosen at random but instead are selected so that an attribute is over- or underrepresented—which leads to a confound. For example, say you wanted to know the average heights of male and females, and you went to shopping malls to measure people. What if you measured the males outside a toy store (and so were likely to be measuring little boys), but measured the women outside a fashion outlet for tall people (and so were likely to find especially tall women)? Or, what if the words in the living/nonliving group in Table 2.3 were interesting words such as centipede and boomerang, and the words in the t/non-t group were bland words such as toe and broom? Or, perhaps the living/nonliving words were more emotionally charged than the t/non-t words, or were more familiar. What if only language majors were

- **Bias:** An investigator’s previous beliefs or expectations alter how a study is set up or conducted, leading it to come out a certain way.
- **Response bias:** A tendency to respond in a particular way regardless of respondents’ actual knowledge or beliefs.
- **Sampling bias:** A bias that occurs when the participants or items are not chosen at random, but instead are chosen so that one attribute is over- or under-represented.

tested, or only people who read a lot and have terrific vocabularies? Could we assume that all people would respond the same way? Take another look at Table 2.3; can you spot any potential sampling bias?

Sampling bias isn't just something that sometimes spoils otherwise good studies. Do you remember the U.S. Presidential election of 2000? Albert Gore and George W. Bush were in a dead heat, and the election came down to the tally in a few counties in Florida. Based on surveys of voters exiting their polling places, the TV commentators predicted that Gore would be the winner. What led them astray? Sampling bias. The news organizations that conducted the surveys did not ask absentee voters how they cast their ballots. In such a close election, this was an important factor because the absentee voters included many members of the armed services, who tend to be Republicans. Thus, sampling only from those who voted on election day produced a biased view of how the entire population voted—and the TV commentators had to eat their words.

### Experimenter Expectancy Effects: Making It Happen

Clever Hans, a horse that lived in Germany in the early 1890s, apparently could add (Rosenthal, 1976). When a questioner (one of several) called out two numbers to add, for example, “6 plus 4,” Hans would tap out the correct answer with his hoof. Was Hans a genius horse? Was he psychic? No. Despite appearances, Hans wasn't really adding. He seemed to be able to add, and even to spell out words (with one additional tap for each letter in the alphabet), but he responded only if his questioner stood in his line of sight and knew the answer. The questioner, who expected Hans to begin tapping, always looked at Hans's feet right after asking the question—thereby cuing Hans to start tapping. When Hans had tapped out the right number, the questioner always looked up—cuing Hans to stop tapping. Although, in fact, Hans could neither add nor spell, he was a pretty bright horse: He was not trained to do this; he “figured it out” on his own.

The cues offered by Hans's questioners were completely unintentional; they had no wish to mislead (and, in fact, some of them were probably doubters). But unintentional cues such as these lead to **experimenter expectancy effects**, which occur when an investigator's expectations lead him or her (consciously or unconsciously) to treat participants in a way that encourages them to produce the expected results. Such effects can occur in all types of research, from experiments to surveys—in all cases, the investigator can provide cues that influence how participants behave.

At least for experiments, it's clear how to guarantee that experimenter expectancy effects won't occur: In a **double-blind design** not only is the participant “blind” to (unaware of) the predictions of the study and hence is unable consciously or unconsciously to serve up the expected results, but also the experimenter is “blind” to the predictions and thus unable to induce the expected results. What would have happened if a questioner of Clever Hans had not known the answer to the question?

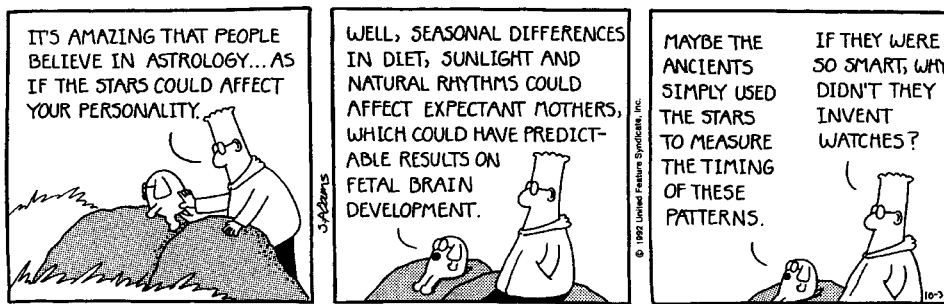
### Psychology and Pseudopsychology: What's Flaky and What Isn't?

Are you a fire sign? Do you believe that your Zodiac sign matters? So many people apparently do that the home page for the *Yahoo!* site on the World Wide Web will automatically provide your daily horoscope. But astrology—along with palm reading and tea-leaf reading, and all their relatives—is not a branch of psychology; it is pseudopsychology. **Pseudopsychology** is superstition or unsupported

● **Experimenter expectancy effects:** Effects that occur when an investigator's expectations lead him or her (consciously or unconsciously) to treat participants in a way that encourages them to produce the expected results.

● **Double-blind design:** The participant is “blind” to (unaware of) the predictions of the study (and so cannot consciously or unconsciously produce the predicted results), and the experimenter is “blind” to the condition assigned to the participant (and so experimenter expectancy effects cannot produce the predicted results).

● **Pseudopsychology:** Theories or statements that at first glance look like psychology, but are in fact superstition or unsupported opinion pretending to be science.



Dogbert (Dilbert's dog) is thinking scientifically about astrology. He proposes a relationship among seasonal differences in diet, sunlight, and other factors and personality characteristics. These variables can be quantified and their relationships tested. If these hypotheses are not supported by the data but Dogbert believes in astrology nevertheless, he's crossed the line into pseudopsychology.

opinion pretending to be science. Pseudopsychology is not just “bad psychology,” which rests on poorly documented observations or badly designed studies and, therefore, has questionable foundations. Pseudopsychology is not psychology at all. It may look and sound like psychology, but it is not science.

Appearances can be misleading. Consider extrasensory perception (ESP). Is this pseudopsychology? ESP refers to a collection of mental abilities that do not rely on the ordinary senses or abilities. Telepathy, for instance, is the ability to read minds. This sounds not only wonderful but magical. No wonder people are fascinated by the possibility that they, too, may have latent, untapped, extraordinary abilities. The evidence that such abilities really exist is shaky, as discussed in Chapter 4. But the mere fact that many experiments on ESP have come up empty does not mean that the experiments themselves are bad or “unscientific.” One can conduct a perfectly good experiment, guarding against confounds, bias, and expectancy effects, even on ESP. Such research is not necessarily pseudopsychology.

Let's say you want to study telepathy. You might arrange to test pairs of participants, with one member of each pair acting as “sender” and the other as “receiver.” Both the sender and receiver would look at hands of playing cards that contained the same four cards. The sender would focus on one card (say, an ace), and would “send” the receiver a mental image of the chosen card. The receiver's job would be to guess which card the sender is seeing. By chance alone, with only four cards to choose from, the receiver would guess right about 25% of the time. So the question is, can the receiver do better than mere guesswork? In this study, you would measure the percentage of times the receiver picks the right card, and compare this to what you would expect from guessing alone.

But wait! What if the sender, like the questioners of Clever Hans, provided visible cues (accidentally or on purpose) that have nothing to do with ESP, perhaps smiling when “sending” an ace, grimacing when “sending” a two. A better experiment would have sender and receiver in different rooms (or better yet, have one on the ISS and another here on earth), thus controlling for such possible confounds. Furthermore, what if people have an unconscious bias to prefer red over black cards, which leads both sender and receiver to select them more often than would be dictated by chance? This difficulty can be countered by including a control condition, in which a receiver guesses cards when the sender is not actually sending. Such guesses will reveal response biases (such as a preference for red cards), which exist independently of messages sent via ESP.

Whether ESP can be considered a valid, reliable phenomenon will depend on the results of such studies. If they conclusively show that there is nothing to it, then people who claim to have ESP or to understand it will be trying to sell a bill of goods—and will be engaging in pseudopsychology. But as long as proper studies are under way, we cannot dismiss them as pseudopsychology.



# Looking at Levels

## Who Will Be a Good Leader?

The ISS will be under the command of a single person, who will have to manage people from various nationalities with different backgrounds and skills. Can psychology offer any insights into the characteristics of a good leader? To answer this question, Simon Taggar, Rick Hackett, and Sudhir Saha (1999) studied 94 teams (each with 5 or 6 undergraduates, who had been together for 13 weeks). They asked the students to complete personality tests and to rate each other for leadership ability. The researchers also obtained a measure of general intelligence. The various test scores were analyzed using a type of correlation, seeing which were most strongly related to the leadership ratings. The most important predictor of who would emerge as a leader turned out to be the level of intelligence, the more the better. But this was not the sole contributor; those perceived to be high in “leadership” were also more likely to be conscientious, extraverted, and emotionally stable. Moreover, the most effective teams were those in which more of the members scored high on such “leadership” characteristics. An effective leader has good followers.

Consider this finding from the levels of analysis perspective. First, the best teams were not simply those that had good leaders. They also needed team members who shared the characteristics of good leaders (level of the group). Second, the characteristics of good leaders included personality characteristics such as being conscientious and extraverted (level of the person). And third, intelligence, which depends on memory and brain processing speed (level of the brain), also proved important for good leaders. As usual, events at the different levels interacted: Good teamwork (level of the group) emerged when team members shared personality characteristics (level of the person) and had higher intelligence (level of the brain).

### TEST YOURSELF!

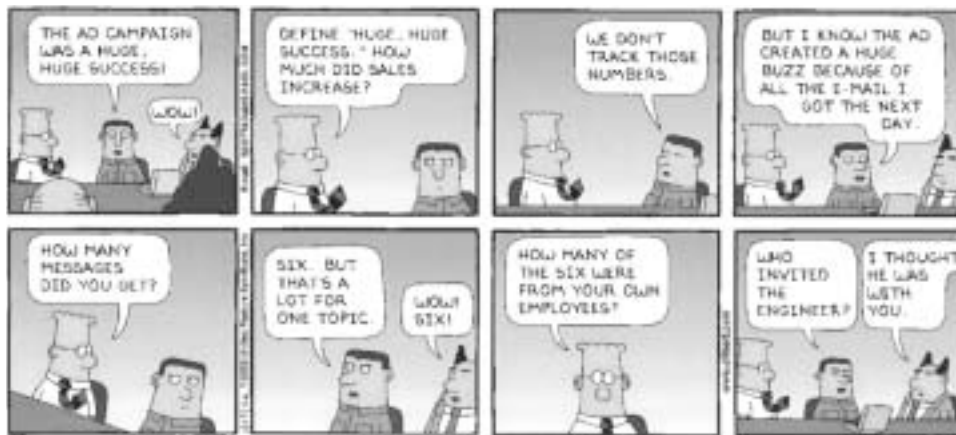
1. What are the key ingredients of experimental research?
2. What is correlational research?
3. What types of descriptive studies do psychologists conduct?
4. Which methodological aspects of studies should you always critically evaluate?

## Statistics: Measuring Reality

Like Rome or the pyramids, the ISS won't be built in a day. This goes not only for the physical structure, but also for how its members will function as a team. Only over time, largely by evaluating current practices and trying to improve them, will researchers discover how best to help astronauts work together smoothly. Over time, researchers will collect various measures of mental processes and performance, such as the rate of human error when operating machines, levels of stress, sleep quality, and memory ability. **Statistics** are numbers that summarize or indicate differences or patterns of differences in measurements. Statistics from the crew of the ISS will be used to guide planners of future missions to outer space, such as the manned mission to Mars.

Mark Twain, borrowing a line from Benjamin Disraeli (Best, 2001), once said that there are three kinds of lies: “Lies, damn lies, and statistics.” The point is that statistics can be used to obscure the facts as easily as to illuminate them. For instance, although the divorce rate in the United States is about 50%, this does not necessarily mean that out of 10 couples only 5 will stay married. If 3 of those 10

● **Statistics:** Numbers that summarize or indicate differences or patterns of differences in measurements.



In trying to assess the success of his company's new ad campaign, Dilbert is searching for statistical evidence of success, rather than relying on people's intuitions.

couples divorce and remarry, and all 3 of those second marriages end in divorce, that makes 6 divorces out of 13 marriages; and if one of those ex-partners remarries a third time and divorces again, we now have 14 marriages and 7 divorces: this is a 50% divorce rate, even though 7 of the original 10 couples stayed married from the start. To understand and evaluate reports of psychological research, be they surveys in newspapers or television or formal research reports in scientific publications, you need to know a few basics about statistics.

## Descriptive Statistics

There are two major types of statistics: One type describes or summarizes data, whereas the other indicates which differences or patterns in the data are worthy of attention. This is the distinction between descriptive statistics and inferential statistics. **Descriptive statistics** are concise ways of summarizing properties of sets of numbers. You're already familiar with such statistics: They are what you see plotted in bar graphs and pie charts, and presented in tables. But descriptive statistics are not limited to figures and tables. For example, in financial news, the Dow Jones Industrial Average is a descriptive statistic, as is the unemployment rate.

You already know a lot about descriptive statistics, but you may not be aware you know it—and you may not be familiar with the technical vocabulary scientists use to discuss such statistics. This section provides a review of the essential points of descriptive statistics.

### Data

As we noted earlier, data are numerical measurements or careful observations of a phenomenon. In other words, in an experiment or quasi-experiment, data are the values of the dependent variable as it varies. Examples of dependent variables used in psychological research are response time (how fast it takes to press a button after perceiving a stimulus), scores on an intelligence test, and ratings of fatigue or the severity of depression.

To understand properties of data, let's consider an example. Astronauts face long hours and grueling work as a normal part of their job and might put to good use a safe medication that could boost memory (particularly if it countered the effects of losing sleep). But before we would recommend taking such a drug, we would want to know whether it really is more effective than a **placebo**, a medically inactive substance, such as a sugar pill. If the drug works as promised, it would be

- **Descriptive statistics:** Concise ways of summarizing properties of sets of numbers.
- **Placebo:** A medically inactive substance that is presented as though it has medicinal effects.

**TABLE 2.4** Fictional Participant Data From Drug and Placebo Conditions

Number of Words Remembered	
Placebo	Memory drug
15	27
12	34
18	21
21	17
22	31
38	47
28	31
15	23
14	40
17	19

in great demand—for instance, by language-learning schools, Wall Street firms, and countless students. To test whether the drug is effective, you ask people to learn a set of words either after taking the drug or, on another day, after taking a placebo. You are interested in whether the participants can later recall more words if they’ve taken your drug than if they’ve taken the placebo; the condition, drug versus placebo, is the independent variable, and the number of words recalled is the dependent variable. Half of the participants get the drug first, and half get an identical-looking and -tasting placebo first. You’ve put the pills in coded envelopes so that your assistant doesn’t know when she’s giving the drug versus the placebo (nor do the participants because you’ve used a double-blind procedure). The comparison would be expressed as the number of words remembered following the drug minus the number following the placebo. The data (scores) from 10 participants are shown in Table 2.4.

### Frequency Distributions

Frequency distributions indicate the number of each type of case that was observed in a set of data. For example, one frequency distribution would indicate how many participants recalled 0 words, 1 word, 2 words, and so on, up to the total of 20 possible words, after taking the drug versus after taking the placebo. Another example of a frequency distribution would be one that indicated the number of men versus women in each state who favor building the ISS. If you worked for a company that manufactured key components of the ISS (earning a hefty portion of that estimated \$90 billion price tag), you could use such data to decide how to spend your advertising budget to promote the ISS: You could target the regions of the country where people were skeptical about the project, or you could write your ads to appeal to men or to women (or both, as the polling data indicated).

### Measures of Central Tendency

When individual measurements are directly presented, they are considered **raw data**; Table 2.4 presents raw data. Descriptive statistics are used to summarize characteristics of a set of such data. Transforming raw data into statistical terms makes the data useful, allowing you to discover and illustrate the relationships among the values or scores. One important type of descriptive statistic is the **central tendency** of the data: the clustering of the most characteristic values, or scores, for a particular group. Central tendency can be expressed three ways. The most common, and probably the one with which you are most familiar, is the arithmetic average, or **mean**, of the scores or values. You calculate a mean by adding up the values in the set of measurements, then dividing that sum by the total number of entries you summed. In Table 2.4, the mean for the placebo condition is 20 words remembered, and the mean for the drug condition is 29 words remembered.

- **Raw data:** Individual measurements, taken directly from the phenomenon.
- **Central tendency:** The clustering of the most characteristic values, or scores, for a particular group.
- **Mean:** The arithmetic average.
- **Median:** The score that is the midpoint of the values for the group; half the values fall above the median, and half fall below the median.
- **Mode:** the value that appears most frequently in the set of data.
- **Normal distribution:** The familiar bell-shaped curve, in which most values fall in the midrange of the scale and scores are increasingly less frequent as they taper off symmetrically toward the extremes.

A second way to specify central tendency is the **median**, which is the score that is the midpoint of the values for the group; half the values fall above the median, and half fall below the median. It is easier to find the median if the data are arranged in order, as shown in Table 2.5. The median in the placebo condition is 17.5, halfway between 17 and 18, the fifth and sixth ordered scores. In the drug condition, the median is 29.0, halfway between the fifth and sixth ordered scores of 27 and 31.

A third measure of central tendency is the **mode**, the value that appears most frequently in the set of measurements. The mode can be any value, from the highest to the lowest. The mode in Table 2.5 is 15 for the placebo condition and 31 for the drug condition.

The mean is the measure of central tendency that is most sensitive to extreme values or scores; if you have a few values at the extreme end of the scale, the mean would change much more than would the median (which often will not change at all). The mode does not generally change in response to an extreme score. For example, if you changed the last score in Table 2.5 in the placebo condition from 38 to 100, the mean would change from 20 to 26.2, but the median and mode would remain the same. When a set of data has many scores near one extreme value and away from the center, it is said to have a *skewed distribution*. When a set of data has a skewed distribution, the median is often a more appropriate measure of central tendency than the mean.

However, the three measures of central tendency generally yield similar results; this is especially likely as the number of observations (data points) becomes larger and the data follow a normal distribution. The **normal distribution** is the familiar bell-shaped curve, in which most values fall in the midrange of the scale, and scores are increasingly less frequent as they taper off symmetrically toward the extremes (see Figure 2.4). Normal distributions occur many places in nature. For example, look at stone stairs in a very old building: You can usually see that they are worn more deeply in the center, and then less so as you move toward the sides. (If the building isn't old enough, you will see the beginnings of a normal curve,

**TABLE 2.5** Fictional Data From Drug and Placebo Conditions, Arranged in Order

Number of Words Remembered	
Placebo condition	Memory drug condition
12	17
14	19
15	21
15	23
17	27
--- <b>Median 17.5</b>	--- <b>Median 29.0</b>
18	31
21	31
22	34
28	40
38	47
Mean* = 20	Mean* = 29
Mode† = 15	Mode† = 31

\*Mean = The arithmetic average.

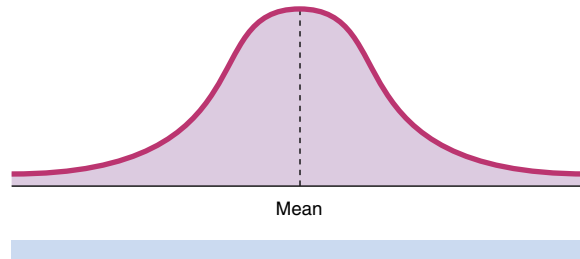
†Mode = Value that appears most frequently.

There are a few very small or very large feet on either end, with most clumping at an intermediate size. The same is true for many psychological qualities, such as scores on intelligence or personality tests.



which over the generations will become deeper in the center until it resembles the shape of the bell curve in Figure 2.4 upside down.)

**FIGURE 2.4** A Normal Distribution



You've probably heard a news announcer say that "Exit polling indicates that it's a dead heat! Smith has 51% and Jones has 49%, with a margin of error of 4%." The margin of error—usually called a *confidence interval*—specifies the range of values within which the mean is likely to fall. The confidence interval is calculated like the standard deviation.

### Measures of Variability

Whereas measures of central tendency convey information about the most common values or scores, measures of variability convey information about the spread of the scores. The **range** is the difference obtained when you subtract the smallest score from the largest, the simplest measure of variability. For the data in Table 2.5, for example, the range of scores in the placebo condition is  $38 - 12$ , or 26. The range of scores for the drug condition is  $47 - 17$ , or 30. But the range does not tell you how variable the scores are in general.

Another method of assessing variability is the **standard deviation**, which is a kind of "average variability" in a set of measurements. In Chapter 9, you will see how important the standard deviation is for understanding intelligence. Here is the key idea: If you take the mean for one group, say the placebo group, and then subtract each of the individual observations (data points) from this mean, you will see how much each score deviates from the mean. These differences are *deviation scores*. But how should you take the average of these scores? Because you've subtracted each observation from the mean, the deviation scores above the mean will equal those below the mean. So, if you took the average of the deviation scores, you would



- **Range:** The difference obtained when you subtract the smallest score from the largest, the simplest measure of variability.
- **Standard deviation:** A kind of "average variability" in a set of measurements, based on squaring differences of each score and the mean and then taking the mean of those squared differences.

get zero; the positive deviation scores would cancel out the negative scores. To get around this problem, you simply square each deviation score (which gets rid of the signs), and then take the mean of these scores (which, as you see in Table 2.6, is 55.6 for the placebo condition). Finally, you need to "unsquare" this mean, by taking its square root. This number is the standard deviation (7.46 for the placebo condition).

For values that are normally distributed, the standard deviation will tell you the percentage of values that fall at different points on the distribution. For instance, about 68% of values fall between one standard deviation below the mean and one standard deviation above the mean. And about 95% of the values fall between two standard deviations below the mean and two standard deviations above the mean.

**TABLE 2.6** Computing the Standard Deviation From the Placebo Condition Data in Table 2.5

Step 1:

(Number of words remembered – Mean)<sup>2</sup> = Deviation score<sup>2</sup>

$$(12 - 20)^2 = -8^2 = 64$$

$$(18 - 20)^2 = -2^2 = 4$$

$$(14 - 20)^2 = -6^2 = 36$$

$$(21 - 20)^2 = 1^2 = 1$$

$$(15 - 20)^2 = -5^2 = 25$$

$$(22 - 20)^2 = 2^2 = 4$$

$$(15 - 20)^2 = -5^2 = 25$$

$$(28 - 20)^2 = 8^2 = 64$$

$$(17 - 20)^2 = -3^2 = 9$$

$$(38 - 20)^2 = 18^2 = 324$$

Step 2: Sum of squares (SS) = Sum of squared deviation scores = 556

Step 3: Variance = SS ÷ Number of deviation scores = 556 ÷ 10 = 55.6

Step 4: Standard deviation = Square root of the variance =  $\sqrt{55.6} = 7.46$

For example, if the placebo condition had a standard deviation of 7.46 words (for simplicity's sake, we'll round this down to 7 words) and a mean of 20 words, then roughly 68% of the participants in this condition will remember somewhere between 13 and 27 words ( $20 - 7$  to  $20 + 7$ ). At two standard deviations from the mean, roughly 95% of participants will remember between about 6 and 34 words.

### Relative Standing

Sometimes you want to know where a particular score stands relative to other scores. For example, college admissions officers want to know how an applicant's SAT scores stand relative to other applicants' scores. One way to convey this information is in terms of measures of variability. You could specify how many standard deviations a score is from the mean. However, this isn't very useful if you are interested in the specific number or percentage of other cases that fall above or below a particular one. Another way of conveying information about a value relative to other values in a set of measurements is to use a **percentile rank**: the percentage of data that have values at or below a particular value. A value converted to a percentile rank of 50, for example, instantly tells you that 50% of the values fall at or below that particular score; the median is a percentile rank of 50. Quartiles are percentile ranks that divide the group into fourths (25th, 50th, 75th, and 100th percentiles); a score that is at the third quartile signifies that 75% of the group falls at or below that score. Deciles are percentile ranks that divide the group into tenths; a score at the sixth decile indicates that 60% of the scores fall at or below that value.

### Inferential Statistics

**Inferential statistics** are the results of tests that reveal whether differences or patterns in measurements reflect true differences or just chance variations. For instance, if you toss a coin 10 times and it lands heads up 7 times, instead of the 5 you would expect purely by chance, does this mean that it is a "trick coin" or an

● **Percentile rank:** The percentage of data that have values at or below a particular value.

● **Inferential statistics:** The results of tests that reveal whether differences or patterns in measurements reflect true differences or patterns versus just chance variations.

edge is worn away, or could this outcome also arise from just chance? Inferential statistics seek to address this question of whether patterns in a set of data are random, or whether they reflect a true underlying phenomenon. A correlation is an example of inferential statistics; if the correlation is high enough (we will discuss what “high enough” means later), it tells you that the scores on one variable do in fact vary systematically with the scores on another variable.

### Correlation: The Relationship Between Two Variables

We’ve seen that a correlation is not about the central tendency and variability of a set of scores, but instead indicates whether two variables are related to each other. Is a change in one variable accompanied by a change in another? To think about how the correlation value is calculated, go back to the idea of a standard deviation. But now, instead of computing the deviations relative to the mean of all the numbers, imagine that you have a line fitting through the data (see Figure 2.3). The closer the data points hug the line, the higher the absolute value of correlation (that is, the higher the number—ignoring whether it is positive or negative). The *method of least squares* is a way to fit a line through a cloud of points. Again, squared numbers are used to eliminate the signs of the difference values. This method positions the line to minimize the square of the vertical distance of each point from the line.

Correlations and other types of inferential statistics may or may not be **statistically significant**. What does “significant” mean? In statistics, it does not mean “important.” Rather, it means that the measured relationship is not simply due to chance. If you correlate any two randomly selected sets of measurements, it is likely that the correlation will not be precisely zero. Say you correlated visual acuity with height and found a correlation of  $-.12$ . Should you pay attention to this correlation, developing a grand theory to explain it? The size of a correlation needed for statistical significance—to be taken as more than just chance variation—depends on the number of pairs of values analyzed (each represented by a point in Figure 2.3). As a general rule, the more observations considered when computing the correlation, the smaller the correlation value needs to be to achieve statistical significance. Why? Imagine that you were randomly throwing darts into a rectangular corkboard on the wall. It is possible that the first few on the left would be lower than those on the right. But as you tossed more and more darts, those initial quirks would be balanced out by quirks later in the process—so after 100 darts, there would no longer be any discernible pattern. In addition, the more observations you have, the less influence extreme values will have.

Statistical significance is expressed in terms of the probability ( $p$ ) that a value (such as the size of a correlation) could be due to chance. For example,  $p < .05$  means that the probability that the result was simply due to chance is less than 5 in 100;  $p < .001$  means that the probability that it was due to chance is less than 1 in 1,000. Usually, any value with  $p < .05$  or smaller is considered statistically significant—not likely to be a result simply of chance variation in the data. You can look up a correlation value in a table to determine its significance, but most computer programs that compute correlation do this for you automatically.

### Samples and Populations

Back to the memory-enhancing drug study. Did the drug work? If you had measured every person on the planet, all you would need to do is look at the descriptive statistics. Either the drug resulted in more learning than the placebo, or it didn’t. But such all-inclusive testing just isn’t practical. Virtually all research in

● **Statistical significance:** The measured relationship is not simply due to chance.

psychology relies on studying data from a **sample**—a group drawn from the population at large—and the goal is to generalize from the findings with the sample to the larger **population**, the group from which the sample is drawn. Inferential statistics let you infer that the difference found between your samples does in fact reflect a difference in the corresponding populations.

Here's a simple example: You are impressed by astronaut Sally Ride's memory, and want to know whether astronauts in general tend to have better memories than the population as a whole (you theorize that to be a successful astronaut you need to remember many facts and procedures). To find out, you send memory tests to current and former astronauts all over the world and include stamped, self-addressed return envelopes. No luck—these people are very busy and don't have time to take your test. So you travel to Houston and get permission to visit the astronauts at NASA headquarters. By some miracle, you are actually able to induce 10 astronauts to take your test. You will next need to compare them with 10 non-astronauts selected to be as similar as possible to your sample—same ages, education levels, gender, and even the same level of fitness. If you then compare the two groups, you will probably find a difference in memory performance. But now consider this: If you had data from 20 astronauts and *randomly* assigned them to two groups of 10, you probably would also find a difference in the mean memory scores for these arbitrarily formed groups! This difference would arise because of how you happened to assign the people to groups. No matter how you did it, the groups would probably have different average memory performance. Only if you had a large number of people would assigning them randomly to two equal-sized groups be likely to result in groups that had nearly identical scores. When you have enough data, people who happen to have unusually good or poor memories will be assigned equally often to each group, on average, and thus their disproportionate contributions will cancel out.

When you compare astronauts and non-astronauts, the problem is to know for sure whether any difference between them is “real”—reflecting actual differences between the two classes in general—or is due to sampling error. **Sampling error** produces differences that arise from the luck of the draw, not because two samples are in fact representative of different populations. In this example, if differences in memory scores between astronauts and non-astronauts are due to sampling error, this means that the two groups are not actually different. (There are statistical tests that can indicate whether a difference between two groups is due to sampling error or reflects a real difference, but details about such inferential tests are beyond the scope of this introduction.)

### Meta-Analysis

Science is a community effort. Usually many people are studying the same phenomenon, each one painting additional strokes onto an emerging picture. **Meta-analysis** is a technique that allows researchers to combine results from different studies. This is particularly useful when results have been mixed, with some studies showing an effect and some not. Meta-analysis can determine whether a relationship exists among variables that transcends any one study, a strand that cuts across the entire set of findings.

Sometimes results that are not evident in any individual study become obvious in a meta-analysis. Why? Studies almost always involve observing or testing a sample from the population; if a sample is relatively small, the luck of the draw could obscure an overall difference that actually exists in the population. For example,

- **Sample:** A group from which one obtains measures or observations that is drawn from a larger population.
- **Population:** The entire set of relevant people or animals.
- **Sampling error:** Non-random sampling from a population, which produces differences that arise from the luck of the draw, not because two samples are in fact representative of different populations.
- **Meta-analysis:** A statistical technique that allows researchers to combine results from different studies, which can determine whether a relationship exists among variables that transcends any one study.

if you stopped the first two males and first two females you saw on the street and measured their heights, the females might actually be taller than the males. The problem of variation in samples is particularly severe when the difference of interest—the effect—is not great. If men averaged 8 feet tall and women 4 feet tall, small samples would not be a problem; you would quickly figure out the usual height difference between men and women. But if men averaged 5 feet 10 inches and women averaged 5 feet 9 inches (and the standard deviation was a few inches), you would need to measure many men and women before you were assured of finding the difference. Meta-analysis is a way of combining the samples from many studies, giving you the ability to detect even subtle differences or relations among variables (Rosenthal, 1991).

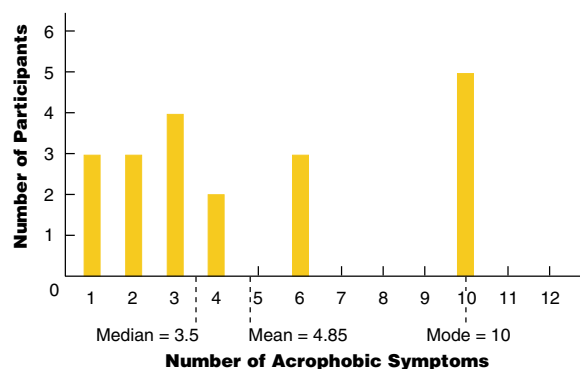
## Lying With Statistics

Statistics can be used or misused. In a famous book entitled *How to Lie with Statistics*, Derrell Huff (1954) demonstrated many ways that people use statistics to distort the pattern of results. Joel Best (2001) has followed in this tradition. Such books play a valuable role in inoculating people against deceptive techniques, and some of their key points are summarized here. Be on the lookout for these manipulations whenever you see statistics.

### Selective Reporting

Because different types of statistics convey different information, the same data can be manipulated to “say” different things. Look at Figure 2.5 and Table 2.7, which present fictitious data for the results of a new type of therapy for people with acrophobia—a fear of heights (obviously a crucial problem to overcome for aspiring astronauts). Before the therapy, participants reported, on average, 9 symptoms of acrophobia; that is, before treatment, the mean number of symptoms was 9. After the therapy, the mean number of symptoms was 4.85, the median was 3.5, and the mode was 10. Proponents of the new therapy make the following claims: On average, symptoms decreased by almost half (based on the mean), and more than 50% of participants had substantial symptom reduction (based on the median). Opponents, however, convey the data differently. A spokeswoman from the pharmaceutical company that manufactures a medication to treat acrophobia makes several counterclaims

**FIGURE 2.5** Number of Symptoms After Therapy for Acrophobia



when she promotes the superiority of her company’s medication: The number of symptoms most frequently reported was 10, which shows that the therapy actually made people more symptomatic (based on the mode). Also, the therapy achieved mixed results, as indicated by the fact that the number of symptoms after treatment ranged from 1 to 10.

As you can see, both supporters and detractors of the new treatment are correct. They are just presenting different aspects of the data. Thus, when hearing or reading about research or survey results, you should ask several questions before taking the results too seriously.

1. What is the distribution of the results? If they are normally distributed, the measures of central tendency will be similar to each other. If they are skewed, the measures of central tendency will convey different information, and the one presented will be the one that conveys the information the reporter wants you to know about. Do the other measures of central tendency paint a different picture of the results?
2. How variable are the data? What does it mean if the results vary a lot rather than a little?

### Lying With Graphs

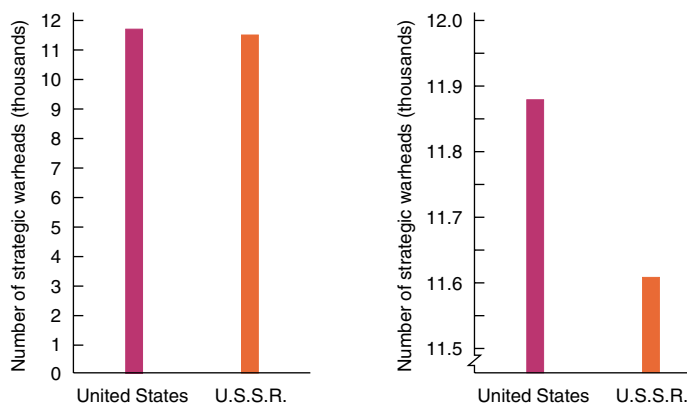
Many results are presented in graph form. Graphs work largely because of a single principle: *More is more* (Kosslyn, 1994a). Larger bars, higher lines, or bigger wedges all stand for greater amounts than do smaller bars, lines, or wedges. Our tendency to see more on the page as standing for more of a substance can lead us astray if graphs are constructed to deceive. Be alert to the following tricks.

**Shortening the Y (Vertical) Axis to Exaggerate a Difference.** As you can see in Figure 2.6, starting the Y axis at a high value and devoting the Y axis to a small part of the scale, as in the right half of Figure 2.6, makes what is in fact a small difference look like a large one. If a difference is statistically significant, it should look that way (and thus shortening the axis may be appropriate). But if it's not, then shortening the axis to exaggerate the difference is deception.

**TABLE 2.7** Fictional Results of Therapy for Acrophobia

Mean number of symptoms after therapy for acrophobia
= Total number of symptoms ÷ Total number of participants
= 1 + 1 + 1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 4
+ 4 + 6 + 6 + 6 + 10 + 10 + 10 + 10 + 10
= 97 ÷ 20 = 4.85
Median = 3.5
Mode = 10

**FIGURE 2.6** Shortening the Y Axis Can Mislead

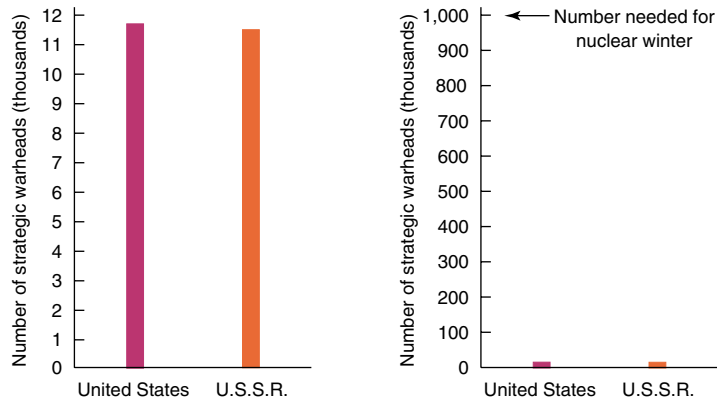


The left panel presents the actual numbers in a neutral way; the right panel exaggerates the difference. (From Kosslyn, 1994a, pp. 209 and 211. Data for illustration on left from Natural Resources Defense Council.)

**Using an Inappropriately Large Range of Values to Minimize a Difference.** The flip side of the coin is illustrated in Figure 2.7, in which a difference is made to appear smaller by using a large range in values on the Y axis.

**FIGURE 2.7** Lengthening the Y Axis Can Mislead

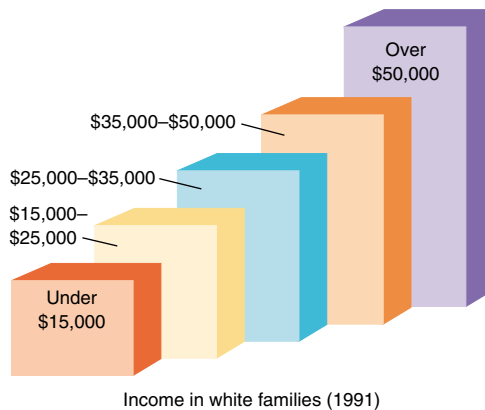
The left panel presents the actual numbers in a neutral way; the right panel minimizes the difference. (From Kosslyn, 1994a, pp. 209 and 211. Data for illustration on left from Natural Resources Defense Council.)



**Using Three-Dimensional Graphics to Exaggerate Size.** As shown in Figure 2.8, a designer can take advantage of our tendency to impose size constancy (see Chapter 4), so that a bar that is farther away will be seen as much larger than a same-size bar that is closer. Even if an actual difference exists, this technique can exaggerate its magnitude.

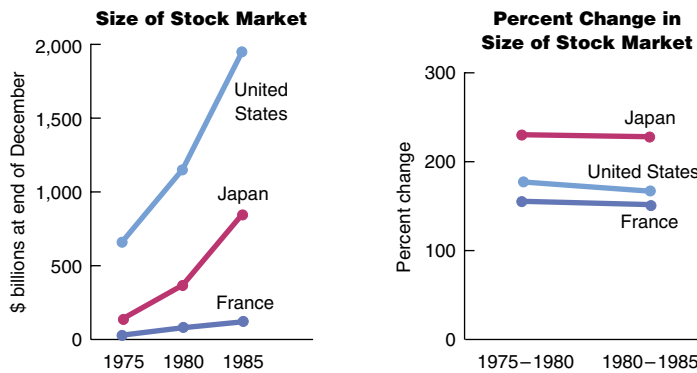
**FIGURE 2.8** Size Constancy Can Exaggerate 3-D Bar Size

Size constancy leads us to see the bars that are farther away as larger than they are, thereby exaggerating a difference. (From Kosslyn, 1994a, p. 227. Data from Hacker, 1992, p. 98, cited in *Newsweek*, 23 March 1992, p. 61.)



**Transforming the Data Before Plotting.** Compare the two panels of Figure 2.9. The first shows the size of the stock market in three countries over 3 years, the second the percentage increase over two 5-year periods. If you saw only the second, you

**FIGURE 2.9** Transforming Data Can Distort the Conclusions



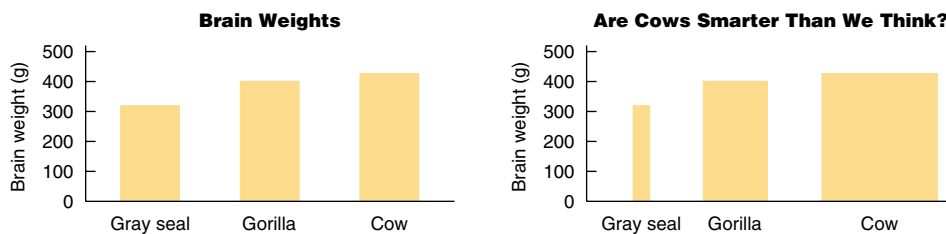
The left panel shows the actual dollar figures, and the right shows the percentage change. Clearly, the message conveyed by the two displays is different. Which one is “more honest” depends on the purpose for which the graph is used. (From Kosslyn, 1994a, p. 219. Data from Morgan Stanley Capital International, cited in *The Economist World Atlas and Almanac*, 1989, p. 90.)

wouldn't realize that the increases in the U.S. stock market were actually much greater than those in Japan. If the user is trying to sell Japanese stocks, you can guess which display will be preferred.

**Changing Width Along With Height.** As shown in Figure 2.10, our visual system does not register height and width separately, but rather we see them simultaneously, as specifying area. So changing the width along with the height gives a much larger impression of amount than is conveyed by changing height alone.

In short, you can see that there is nothing magical or mysterious about statistics or how they are represented visually. Whenever you see a graph in the newspaper, you are seeing statistics; when you hear that a poll is accurate to “plus or minus 3 points,” that's the spread within which the mean is likely to occur if you look at other samples. The crucial ideas are that there are measures of central tendency (mean, median, and mode), measures of variability (such as the range and standard deviation), and statistical tests that tell you the likelihood that a measured difference is due to chance alone. What you've learned here is enough to enable you to read and understand many reports of original research in psychology.

**FIGURE 2.10** Changing Width With Height Exaggerates Size



Expanding the bar width conveys the impression that more quantity is being presented than increasing the height alone signals. (From Kosslyn, 1994a, p. 225. Data from Weisberg, 1980, cited in Chambers et al., 1983, p. 371.)



# Looking at Levels

## Graph Design for the Eye and Mind

Astronauts on the ISS will have to monitor many sources of data, both for the experiments they conduct and to ensure that the space station functions effectively.

Graphs are a way to convey a lot of information without overwhelming the user. But what kind of graph should be used? It depends on what message needs to be conveyed. For example, Jeffrey Zacks and Barbara Tversky (1999) found that bar graphs are better than line graphs when you need to make or illustrate comparisons between discrete data points (such as specific numbers of Democratic versus Republican voters who support the space program), whereas line graphs are better when you need to understand or illustrate trends (such as changes in the numbers of Democratic and Republican supporters in different parts of the United States over time). Bars end at discrete locations, and thus it's easy to compare data points simply by comparing the heights of the bars. In contrast, bars are not as useful for conveying trends because the reader needs mentally to connect the tops of bars, creating a line in order to determine visually whether there is a

trend. Hence, if that's what you want to convey, it's better to give the reader the line in the first place. But if the reader needs to compare discrete data points, a line isn't so good: Now the reader must "mentally break down" the line into specific points, which requires effort (Kosslyn, 1994a).

Think about this finding from the levels perspective: When designing a graph, you should create one that most effectively communicates to other people (level of the group). To do so, you need to respect the way the human perceptual and conceptual systems work (level of the brain). And you need to keep in mind that the best graph type minimizes the effort required of the reader; people will be more likely to understand a graph if they aren't forced to work hard (level of the person). And, of course, events at the different levels interact: If the graph is so hard to read (level of the brain) that people aren't motivated to decipher it (level of the person), it won't end up communicating anything (level of the group).

### TEST YOURSELF!

1. What are descriptive statistics?
2. What are inferential statistics? How do they differ from descriptive statistics?
3. How can people (intentionally or not) lie with statistics?

## How to Think About Research Studies

Large amounts of research have been reported from studies of human performance in space. Unfortunately, many reports of this important research are not as easy to understand as they should be. If a piece of research is going to have an impact (for example, on the design of a future space station or expedition to another planet), it must be read and understood. We have found it useful to approach reading—and writing—research reports armed with what we call the *QALMRI method*. This method is a vehicle for understanding the meaning of a research study in the literature—and for reporting your own research. This method will help you become clear about what question is being asked, how the researchers have tried to answer it, and whether the results really do support the preferred answer (the hypothesis).

### Reading Research Reports: The QALMRI Method

When you read a research report, try to identify the following components.

### **Q Stands for the Question**

All research begins with a question, and the point of the research is to answer it.

The first few paragraphs of the General Introduction should tell the reader what question the article is addressing. In addition, the context provided by the General Introduction's review of previous studies should explain why the question is important, why anybody should care about answering it. In some cases, the question is important for practical reasons, whereas in others it is important as a way to test a theory (and, in some cases, it is important for both practical and theoretical reasons, as in our example with placebo). The General Introduction should provide the general context, explaining the reasons why the question is worthy of consideration.

### **A Stands for Alternatives**

A good report describes at least two possible answers to the question and explains why both are plausible. After describing the question that is being addressed, the General Introduction should explain what alternatives are being considered. When reading the General Introduction, identify the question and then the alternative answers that will be addressed by the study. If the alternatives are not spelled out, try to figure out for yourself what they might be; if the study is simply seeking to confirm a theory's prediction, try to get a sense of whether other theories (or just common sense) would make the same prediction. If all of the theories make the same prediction, it probably isn't worth testing.

### **L Stands for the Logic of the Study**

The goal of the study is to discriminate among the alternatives, and the logic is the general idea behind the study—the way the study will distinguish among the alternatives. The logic is typically explained toward the end of a study introduction and has the following structure: *If* alternative 1 (and not the other alternatives) is correct, *then* when a particular variable is manipulated, the participant's behavior should change in a specific way. For example, the logic of the memory-enhancing study described earlier was: "If the drug enhances memory (and the placebo doesn't), then people should recall more test words after taking the drug than after taking a placebo."

### **M Stands for the Method**

The details of what the researcher did are found in the Method section. The Method section has the following parts:

*Participants:* Look to see how the participants were selected. Are they a representative sample of the population of interest? If a study was conducted to make a recommendation for a particular type of people (such as men and women in their early 20's), then the participants should be as similar to that group as possible. If no particular population is specified, then the sample should be representative of the population in general. If the study involves more than one group, they should be equivalent on important variables, such as age and education. Depending on the study, variables such as the level of depression, experience with medicine, or experience in large, noisy brain-scanning machines can be relevant. Try to think of all possible confounds that could make the groups different in ways that might affect the study's outcome.

*Materials:* If questionnaires are used in the study, they should have been shown to be valid (that is, they should measure what they are supposed to measure). And they should be reliable (that is, they should produce consistent

results). In addition, materials used in different parts of the study should not differ except as required to answer the research question.

*Apparatus:* The apparatus delivers stimuli or defines the experimental situation. If a computer is used, the authors should describe exactly how it presented the stimuli. They also should describe in detail all other physical props they used. Think about how the apparatus looked to the participants and whether it could have distracted them or allowed them to pick up inappropriate cues.

*Procedure:* The procedure is the step-by-step process of what happens in a study. Try to picture yourself in the study. A good procedure should be described so well that you could replicate the study, doing exactly the same thing as the original investigators. Were participants given appropriate instructions (clear, but not leading them on)? Was it clear that the participants did in fact understand the instructions? Could the investigator have unintentionally treated participants in different groups differently?

### **R Stands for the Results**

The outcome of the study is described in detail in the Results section. What happened? First, look for measures of central tendency (means, medians, modes) and some measure of the sampling variability (commonly, standard deviations). The actual results—what the researchers found—are descriptive, and often are presented in a graph or table. Second, not all differences and patterns in the results should be taken seriously; some differences are simply quirks due to chance. Inferential statistics should be reported to indicate which patterns of variation are unlikely to have arisen due to chance. Look for the “p values” that document differences; if the p value is .05 or less, you can be reasonably certain that the difference found in the sample reflects an actual difference in the population as a whole.

### **I is for Inferences**

The payoff of a study is the inferences that can be drawn about the alternative answers to the question being asked, given the results that were obtained. Look to see whether the researchers convincingly answered the question they posed at the outset. The Discussion section usually contains the inferences the authors want to draw from their results. If the study was well designed (the logic sound and the method rigorous), the results should allow you to eliminate at least one of the alternatives, and ideally should be most consistent with only one of the alternatives. At this point, take a step back and think about potential confounds that could have led to the results. Were any alternative explanations not ruled out? For example, perhaps participants in different groups were treated differently by the investigators, or perhaps they were tested at different times of day or at different periods in the semester (closer or farther from anxiety-inducing exams). And consider any loose ends—what else would you want to know about the phenomena?

In sum, the QALMRI method helps you focus on the “big picture”: What a study is about, why it’s important, and what the results actually mean. When you read a study, figure out exactly what question the authors wanted to answer and what alternative answers they’ve considered. Can you think of others? Always be on the lookout for potential alternative explanations, and look for features of the study that limit how well its results can be generalized; for example, can you assume that the results necessarily apply to other ages, races, or cultures? Be sure to read the footnotes. The single most important advice we can give about reading a

study is to be an active reader: Think about what the authors are claiming, and think about whether it makes sense.

## Writing Your Own Research Papers

The same principles apply to writing your own research papers. Write the Introduction so that the reader clearly understands the question you are addressing and why it is important. Your question can be important because it is an extension of previous research (which you summarize in the Introduction), or because you've spotted a hole in the literature and aim to fill it by supplying new information, or because you've identified a variable that might invalidate a previous study (and want to find out whether those researchers did in fact overlook something crucial). When you review previous studies and theories in the published literature, only review those that help you explain why your question is worth considering, that put it in context. Abraham Lincoln was once asked how long a man's legs should be, and replied, "Long enough to reach the ground." The same principle applies to Introductions: Don't include any more or less material than you need to put your question in context.

The Introduction should also explain the alternative possible answers you will consider—including, in most cases, your "favorite" one, which is called "the hypothesis." You need to explain why each alternative is plausible, usually by referring to previously published findings and theories. Finally, the Introduction should end with a clear statement of the logic of your study, the basic idea underlying what you did.

In the Method section, be sure to include enough detail to allow another researcher to repeat exactly what you did. Explain what sort of participants were tested, and how you ensured that participants in different groups were comparable in terms of important variables. In addition, you need to describe the materials in detail, and you also need to describe precisely the apparatus and the procedure.

In the Results section, first present results that bear directly on the question and alternative answers. The results that address the question being asked are most important—even if they are not as striking as some of the other findings. If your Introduction is clear, the reader is focused like a laser beam on the question you are asking and is waiting to find out which alternative answer is supported by the results. Don't keep the reader in suspense; present the results that speak to the question at the outset of the Results section. These results should be measures of central tendency and variability, which are often best presented in a graph; you should also present inferential statistics along with the results, so the reader will know which differences to take seriously. After you present these results, present everything else that you may have found.

Finally, in the Discussion section, return to the question and alternative answers, and discuss exactly what you can infer from your results. Have you shown that some of the alternatives must be discarded? Is only one viable? What should future research focus on to propel the field even further ahead?

When writing a research report, always put yourself in the place of the intelligent reader. If a report has been written clearly, the reader will glide through it effortlessly, understanding what the author intended to convey, why the research was conducted in a particular way, what the discoveries were, and why the report is interesting and important.



## UNDERSTANDING RESEARCH

### When Does Mental Practice Improve Later Performance?

In each chapter, we will examine one study in detail, using the QALMRI method. In this chapter, let's look at *mental practice*, the ability to rehearse an activity mentally, without actually making any movements. Mental practice is particularly interesting in the context of the ISS. Mission planners intend to give astronauts "leisure time," and on earth most astronauts would use at least some of this time to play tennis, golf, or some other sport. Unfortunately, few of these sports can be played on the ISS—balls don't bounce properly in weightless environments, and there really isn't room to run around (which would jeopardize delicate equipment, even if there were room). Does this mean that the astronauts must resign themselves to getting rusty at their favorite sports? Perhaps not. Perhaps they can practice mentally, which would preserve—or even improve—their game. For example, many golfers claim that when they are off the course, they can practice by imagining themselves whacking the ball straight down the fairway or out of the sand trap. Players regularly claim that the mental practice improves their game. Well, maybe. The only way we can find out whether mental practice really works is by conducting a scientific study, and many such studies have been reported. Let's now consider one of them.

**QUESTION:** Can mental practice change subsequent golf putting? Woolfolk, Parrish, and Murphy (1985) asked whether mentally rehearsing golf putts can help as well as hurt subsequent performance.

**ALTERNATIVES:** (1) Mental practice improves putting when participants imagine successfully tapping the ball into the hole, but it actually hurts performance when they imagine tapping the ball so that it misses the hole; (2) Mental practice might always improve putting; (3) It might not have any effect at all.

**LOGIC:** If Alternative 1 is correct and the other alternatives are not, then when people imagine rehearsing the right kind of movements for a successful putt, their performance should later improve—but if they imagine rehearsing the wrong kinds of movements, their performance should actually get worse.

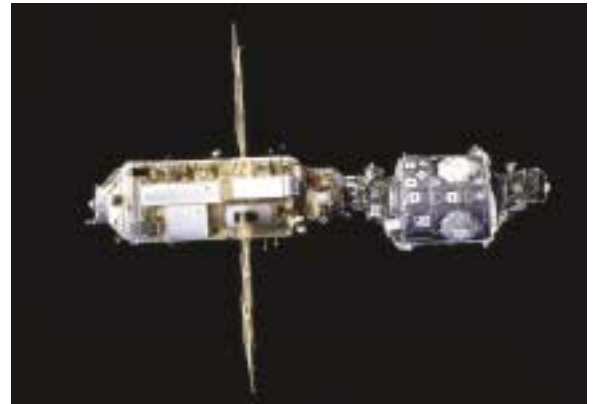
**METHOD:** The researchers first asked 30 college students to putt golf balls into a hole and assessed how well they could do so. After performing 20 putts (from 8.5 feet away), equal numbers of students of comparable skill were randomly assigned to each group. The researchers then gave each group different instructions for mental rehearsal. They asked students in the *positive imagery group* to imagine making a "gentle but firm backswing," and seeing the ball "rolling, rolling, right into the cup" (p. 338). Students in the *negative imagery group* received the identical instructions but were told to imagine the ball "rolling, rolling, toward the cup, but at the last second narrowly missing." Finally, they asked students in the *control group* to imagine putting, with no specific instructions about how to imagine the ball. The students then imagined practicing, following the instructions given to their group. After this, the researchers again asked the students actually to putt and again assessed how well they could do.

**RESULTS:** Students in the positive imagery group performed 30.4% better after mental practice than they had when tested initially. In contrast, students in the negative imagery group actually got worse, scoring 21.2% poorer than they had earlier. Finally, students in the control group improved a bit (9.9%).

## UNDERSTANDING RESEARCH *(continued)*

**INFERENCES:** The authors conclude that mental practice depends on the specific movements you imagine. If the movements are appropriate, mental practice will help later performance—but if the movements are not appropriate, mental practice will actually hurt later performance. The students in the control group apparently often imagined putting correctly, but not as often (or as effectively) as the students in the positive imagery group. Many other studies have found that mental practice improves subsequent performance (Doheny, 1993; Driskell et al., 1994; Druckman & Swets, 1988; Prather, 1973; Vieilledent et al., in press; White & Hardy, 1995), and the present results begin to suggest why it might work.

You might wonder, however, whether the results occurred not because of differences in the images, but because the students in the positive imagery group were more relaxed than those in the negative imagery group. Or perhaps the students in the negative group found it frustrating to keep missing the hole, and thus stopped practicing altogether. Or perhaps at the time of the second actual testing, students in the negative imagery group thought the experimenter expected poorer performance, and so “threw the game” and performed more poorly than they could have. Each of these alternative explanations can be tested.



One glance of the International Space Station from the outside makes it clear why astronauts would get a lot out of mentally practicing sports—there just isn't much room on board for the real thing.



## Looking at Levels

### Imitation Is the Sincerest Form of Flattery

Imagine what life would be like if every astronaut had to learn the job solely by trial-and-error. Fortunately, we can often learn by observing and imitating others. In fact, mental practice appears to rely on many of the same neural mechanisms as does imitation—you are mentally imitating what you “see” in a mental image. This idea is plausible because many researchers have found that a set of the same brain areas is activated when people perform an action, watch somebody else perform an action, or simply imagine performing the action (e.g., Decety, 2001; Grèzes & Decety, 2001). Decety and colleagues (2002), reflecting on such findings, raise a conundrum: If the same brain areas are used in these different cases, how do we know when we've performed an action versus only watched someone else perform it? Their hypothesis was that certain brain areas keep track

of exactly this distinction, whether you or someone else has performed the action. To test this hypothesis, researchers monitored participants' brain activation while they either imitated the experimenter's hand movements as he manipulated small objects, or themselves made such movements and watched the experimenter imitate them (the participants also took part in three other conditions that did not involve imitation, which allowed the researchers to zero in on which brain areas are involved in imitation per se). The important finding was that the left portion of one brain region (the bottom part of the parietal lobe, as is explained in the following chapter) was more active when the participants imitated the experimenter, but the right portion of this same region was more active when the participants watched the experimenter imitate their actions. This area of the brain seems to play a critical role in distinguishing whether you produce an action or somebody else does.

Consider this finding from the levels of analysis perspective. First, imitation is by definition a social affair: You



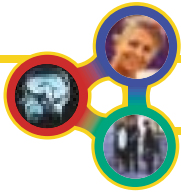
## LOOKING *at* LEVELS (continued)

need somebody else to imitate, or to imitate you (level of the group). Second, the brain registers what other people do similarly to what you yourself do—which allows others' actions to guide you in the future (level of the brain). However, the brain does not respond identically in the two conditions; rather, subtle differences distinguish between your making actions versus your watching actions that someone else makes. Third, you need to be motivated to watch someone else, or you won't learn by imitation (level of the person). Events at the three levels interact: If nobody was available to show you how to perform an action (such as operate a space shuttle or play golf) then, even if you were

motivated to learn, you wouldn't have the chance—and your brain would not have the opportunity to respond in a certain way. But once your brain has had such good fortune, this opens up more opportunities for choosing actions (level of the person), which in turn will affect others (perhaps providing them with the opportunity to learn from you).

### TEST YOURSELF!

1. What does QALMRI stand for? How can you use the QALMRI method when you read research reports?
2. How can you use the QALMRI method when you write your own research papers?



## CONSOLIDATE!

### The Scientific Method: Designed to Be Valid

- The science of psychology relies on the scientific method, which involves specifying a problem, systematically observing events, forming a hypothesis of the relation between variables, collecting new observations to test the hypothesis, using such data to formulate and support a theory, and testing the theory.

**THINK IT THROUGH** Think of 5 questions about the way being cooped up with 7 other people for months on end could change relationships (imagine that there are 4 men and 3 women in total). Can each one of your questions be answered using the scientific method? Why or why not? What are the limits of the scientific method for studying psychology? Are there any? If you think there are such limits, what other methods could you use to study such aspects of mental processes and behavior?

### The Psychologist's Toolbox: Techniques of Scientific Research

- Psychologists test hypotheses and look for relations among variables using a variety of tools, including experiments, quasi-experiments, correlational studies, naturalistic observation, case studies, and surveys.
- In an experiment, the effect of manipulating one or more independent variables on the value of a dependent variable is measured, and participants are assigned randomly to groups.
- Quasi-experiments are like experiments but participants are not assigned to groups randomly.
- In correlational studies, the relationship between the values of pairs of variables is assessed, showing how the values of one go up or down as the values of the other increase (but not showing that changes in the values of one variable *cause* changes in the other).
- Naturalistic observation involves careful observation and documentation of events.
- Case studies are detailed investigations of a single instance of a situation (the detailed exploration of an astronaut's training would be a case study).
- In surveys participants are asked to answer sets of specific questions.
- When reading reports of studies, you should be alert for the following: (1) evidence that the data are reliable, (2) evidence that the data are valid, (3) possible contamination from confounding variables, (4) biases, including the tendency to respond in particular ways to everything (response bias) and the nonrandom selection of participants or experimental materials (sampling bias), and (5) experimenter expectancy effects
- Pseudopsychology differs from psychology not necessarily in its content, but in how it is supported by data.

**THINK IT THROUGH** If you wanted to know whether Sally Ride’s upbringing played a crucial role in leading her to become an astronaut, how would you go about studying this? Don’t assume that it has to be a case study. Which specific questions would you ask? What are the best methods for answering them?

What characteristics and qualities do you think an astronaut should have? Do you think psychologists should prevent anyone from entering astronaut training who does not have these characteristics and qualities? Why and why not?

## Statistics: Measuring Reality

- Descriptive statistics characterize observations by specifying measures of central tendency and variability.
- Measures of central tendency include the mean (which is the arithmetic average), median (which is the numbers for which half the other numbers are higher and half are lower), and mode (which is the value at which the most observations occur).
- Measures of variability include the range (which is the difference between the highest and lowest score) and standard deviation (which is a measure of “average spread” from the mean).
- Different descriptive statistics indicate the frequency of different scores and the standing of any one score relative to the others (for example, in terms of quartiles or deciles).
- Inferential statistics tell you which differences among values or patterns (such as increasing or decreasing trends) in the data should be taken seriously. Inferential statistics rely on assigning a probability that a difference or pattern could have arisen purely due to chance. Generally speaking, if that probability is less than 5 times in 100, the result is considered “statistically significant.”
- A correlation indicates whether one set of measurements tends to vary along with another set.
- A meta-analysis identifies trends or patterns that are present across many studies.
- Inferential statistics can be used deceptively, largely because of selective reporting.
- Graphs can be constructed to bias the interpretation of the reader, either appropriately emphasizing the actual results (statistically significant differences or patterns in the data) or inappropriately emphasizing non-significant results.

**THINK IT THROUGH** Aspiring astronauts take a variety of psychological tests. If the values of the mean, median, and mode are not the same, which one should you take most seriously? To what extent does this depend on the purposes to which you will put these data, and to what extent should your confidence reflect properties of the measures themselves? Would it matter how many observations you have?

People sometimes argue that “garbage in, garbage out”: If the data you begin with are no good, you

won’t be able to use them to draw inferences. This observation is sometimes applied to meta-analyses, which often include studies that have flaws. To what extent do you think it applies? Should all meta-analyses only include “perfect” studies? What if the flawed studies are flawed in different ways, so that the flaws are not correlated with the outcomes?

## How to Think About Research Studies

One way to think about the relation between theory and data relies on the QALMRI method.

- **Q** stands for the **question**, what the study is about and why it is important.
- **A** stands for **alternative** answers to that question, which the study is designed to discriminate among.
- **L** stands for the **logic** of the study, the basic idea that will allow the researchers to discriminate among the alternatives.
- **M** stands for the **method**, the details of exactly what was done in the study.
- **R** stands for the **results**, which include both descriptive and inferential statistics.
- **I** stands for **inferences** that can be drawn from the results, indicating which alternatives can be eliminated and which receive support.

**THINK IT THROUGH** Pick your favorite hobby and design an experiment to discover whether mental practice could improve your performance. Use the QALMRI framework to describe the study.

## Key Terms

- bias, p. 41
- case study, p. 37
- central tendency, p. 45
- confound, p. 34
- control condition, p. 35
- control group, p. 34
- correlation, p. 36
- data, p. 31
- dependent variable, p. 33
- descriptive statistics, p. 44
- double-blind design, p. 42
- effect, p. 33
- experimenter expectancy effects, p. 42
- hypothesis, p. 31
- independent variable, p. 33
- inferential statistics, p. 48
- mean, p. 45
- median, p. 45
- meta-analysis, p. 49
- mode, p. 45
- normal distribution, p. 46
- operational definition, p. 31
- percentile rank, p. 48
- placebo, p. 45
- population, p. 49
- predictions, p. 32
- pseudopsychology, p. 42
- random assignment, p. 35
- range, p. 46
- raw data, p. 45
- reliability, p. 38
- replicated, p. 31
- response bias, p. 41
- sample, p. 49
- sampling bias, p. 41
- sampling error, p. 49
- scientific method, p. 30
- standard deviation, p. 46
- statistical significance, p. 48
- statistics, p. 44
- survey, p. 38
- theory, p. 32
- validity, p. 39
- variable, p. 31