3

IN A NUTSHELL

An Overview of Psychological Research Methods

A TALE OF TWO VALIDITIES

So far in this text, you've learned that scientists make decisions about what is true very differently than non-scientists. Relative to laypeople, scientists place more emphasis on data and logic, for example, and they place less emphasis on authority and intuition. You should also recall that *psychological* scientists face practical and ethical challenges that physical scientists such as chemists and astronomers rarely have to consider. No one worries about whether it is ethical to figure out why iron oxidizes faster than copper. Likewise, no one has to get informed consent from distant stars to track their movements. Like human thoughts and feelings, black holes are notoriously hard to observe directly, but no one has to get their permission to do the observing. Internal review boards (IRBs) exist because psychological scientists must think carefully about participants' rights (and their own responsibilities) before conducting scientific research on people. But before you get permission from an IRB to conduct a study and thus become a card-carrying empiricist, you have to think pretty hard about exactly what kind of observations you plan to make in the first place. If the folks who serve on an IRB are doing their job properly, they will always put at least a little weight on the *value* of your research. The **risk-benefit rule** you learned about in Chapter 2 means that if there any risks at all to the participants who take part in research, there must be some potential benefits to society to offset those risks.

There are two ways to ensure such benefits. The first way is to choose a research problem that matters. All else being equal, a study that could help reduce teenage pregnancy rates has more value than a study that could help reduce ice cream spoilage rates (not that there's anything *wrong* with reducing ice cream spoilage). Likewise, a study that could help reduce rates of unnecessary surgery has more potential benefit to society than a study that could help reduce rates of unnecessary paper cuts. Of course, there's often room for debate about exactly what's important. A cure for male pattern baldness probably seems more important to the male, middle-aged authors

of this textbook than it does to you. (Nice haircut, by the way. Neither of us could pull that off.) But almost all psychological scientists agree that social problems such as bullying, clinical depression, obesity, and climate change are important. Further, almost all psychologists agree that more highfalutin topics such as persuasion, attention, and decision-making are important because they can affect anything from whom you choose to marry to why presidential elections turn out the way they do.

But choosing an important research question does not, by itself, guarantee that your research will have value. To be truly valuable, research must also tell us something. Ideally, this will be something pretty specific and robust, about a well-specified research question. To be truly valuable, research must be methodologically rigorous. It must inform. As it turns out, much of the rest of this textbook is about designing and conducting research that is highly informative in just two basic ways. Research is almost always informative because it maximizes **internal validity** (information about what causes what) or because it maximizes **external validity** (information about how well a research finding holds up in the real world). That's right; the rest of this entire book is focused mainly on how to make sure research has internal validity, has external validity, or has both. If you are reading this chapter, there's a good chance your instructor has decided to get you started conducting a research project early enough in your academic term that you can complete the project well before finals week arrives. You can be guaranteed that your instructor will want your project to have plenty of both internal and external validity. The only problem with this hands-on approach is that you won't know most of the details of how to conduct good research until you have completed this book! Just as no one should begin playing a board game without knowing the basic rules of the game, no one should begin conducting a study in psychology without knowing the basic rules of psychological science.

This chapter is an effort to solve this dilemma by introducing you to the two most important rules of research design *before* you conduct any research. By becoming familiar with a couple of key principles before you collect any data, you can maximize the chances that your data will be valuable. In short, then, this chapter is a preview of much of what is to come in the rest of this book. It is a conceptual user's manual for getting a handle on the cardinal principles of conducting psychological research. We hope this chapter also has another kind of value, by the way. Because this chapter introduces readers to a set of basic (and easy to remember) rules for doing good research, completing this chapter should also put you in a good position to evaluate existing research, including the research of those who have years of experience on you. At the risk of straining the metaphor of board games, this chapter can't make you a chess master or a great bridge player in just a few hours. But it can certainly teach you to follow the most important rules of the game of science. Where you go next with some basic rules under your belt is up to you and your instructor. If your instructor does want you to start a research project pretty early in the semester, you're going to want to take a close look at Chapter 4 after you complete this summary chapter.

THREE REQUIREMENTS FOR ESTABLISHING CAUSALITY

Recall that there are only two truly basic principles of good research (maximizing internal validity and maximizing external validity). The first principle has to do with establishing **causality**. With very few exceptions, that is, psychologists collect data with the goal of uncovering the *causes* of human behavior. Recall that *theories* are formal ideas about what causes what. From this perspective, one could argue that the canon of *determinism* discussed earlier in this text trumps all the other scientific canons—because most scientists are obsessed with causes. After all, if you know *why* something is true rather than just knowing *that* it's true, you will have made the world a more predictable and orderly place. Determinism is all about understanding causes. But how do scientists uncover causes?

Covariation

Most researchers who wish to understand causality rely heavily on the logical framework proposed by the British philosopher John Stuart Mill. If we may simplify Mill a bit, he proposed five methods that can be distilled down to three basic requirements for establishing that one thing causes another (Mill, 2002/1863, and see Copi, 1978, for a modern treatment of the five original methods). The first of Mill's requirements, **covariation**, is probably the easiest. For one variable to cause another, Mill argued, changes in one variable must correspond with changes in the other. As an example, many people strongly believe that the hormone testosterone causes aggression. However, a problem with this argument is that there is surprisingly little evidence that increasing a person's testosterone levels increases that person's tendency to behave aggressively. If testosterone levels aren't really *correlated* with aggression, it's pretty hard to argue that testosterone levels cause aggression. On the other hand, recent research suggests that even though testosterone may not automatically foster aggression, high levels of testosterone *are* associated with a desire for competition and social status. That argument seems much safer (Boksem et al., 2013). Along similar lines, because there is a very clear correlation (because there is *covariation*) between biological sex and aggression, researchers spend a lot of time debating exactly which aspects of being male (e.g., hormonal aspects or cultural aspects) are responsible for the fact that men are about ten times more likely than women to commit highly aggressive acts such as murder. If there were no covariation between gender and aggression in the first place, no one would debate the exact sense in which being male causes people to be aggressive.

Across animal species there is a sizable correlation (i.e., strong *covariation*) between total body mass and bone thickness. Elephants have *much* thicker bones—even

relative to their total body size—than do cats or mice (see Figure 3.1). And the largest land-dwelling dinosaurs had bones proportionally thicker than those of elephants. This is partly a requirement of physics. The total weight of any animal increases as a cubed function of its linear size—because the mass of the animal increases as its height *and* width *and* depth increase. Consider a normal and a gigantic man with identical proportions. If the 6-foot tall man weighed 200 lb., his 12-foot-tall proportional equivalent would weigh 1,600 lb.! This is because the giant would be twice as tall *and* twice as wide *and* twice as thick ($200 \times 2 \times 2 \times 2 = 1{,}600$). If elephants didn't have extra stocky bones, then their skinny bones would snap under their own weight. Notice that a predictable exception to this rule is the elephant's wispy tail. If you are wondering what the heck this has to do with psychology, the answer is that a growing field of study in psychology is evolutionary psychology. The fact that animals vary in ways that help them survive and reproduce is consistent with a basic premise of natural selection: As a rule, evolution is usually pretty darn efficient. One reason why some animals have thicker bones than others, then, is because thicker bones are a requirement of survival if you are a large land mammal. One could also generate more subtle predictions. For example, it is well established that controlling for bone thickness, the bones of birds are much lighter than the bones of reptiles. Because birds with very dense bones would have great difficulty flying, this kind of covariation between biological family and bone density is surely no evolutionary accident.

No one can easily make a claim about causation in the absence of covariation. But covariation by itself is not enough to establish causality. Consider divorce and distress. They are certainly correlated. And it might seem obvious that distress causes divorce.

Figure 3.1 An elephant skeleton and a cat skeleton. The cat's weight-bearing bones are proportionally much thinner than those of the elephant. Bone thickness *covaries* with body weight.

Photo of Elephant Skeleton by SkImsta, https://commons.wikimedia.org/wiki/File:Elephant_skeleton.jpg, licensed under CC0 1.0 https://creativecommons.org/publicdomain/zero/1.0/deed.en

Photo of Australian Native Cat by Cliff, https://commons.wikimedia.org/wiki/File:Dasyurus_maculatus_skeleton.jpg, licensed under CC BY 2.0 https://creativecommons.org/licenses/by/2.0/deed.en

People get divorced because they are *unhappy*, don't they? Probably, but divorce itself can also cause many people to *become* distressed (Ambert, 2009). Simple covariation by itself is highly consistent with either temporal order. A nice guy like John Stuart Mill was probably very happily married, but he knew very well that covariation alone is only one piece of the puzzle of understanding causality.

Temporal Sequence

Mill's second requirement is **temporal sequence**. To argue that changes in one variable cause changes in a second, one must be able to show that the changes in the first variable *preceded* the changes in the second. This is not always easy to do. For instance, researchers often measure a wide range of variables *at the same time* to see if different variables covary with one another in ways predicted by a particular theory. With this kind of **passive observational** (i.e., cross-sectional) **research design**, it's often impossible to establish temporal sequence (as would be the case in a cross-sectional survey of divorce and distress). In correlational studies, it's often impossible to know what caused what. In light of this problem, researchers sometimes measure variables over time. In **prospective designs** (e.g., longitudinal studies), researchers measure all of the variables of interest on at least two different occasions. They can thus see if changes in one variable do, in fact, precede changes in a second. Two drawbacks of prospective designs are the closely related facts that prospective research is both expensive and time consuming. For this reason, only a small minority of all studies in psychology are prospective (e.g., longitudinal studies). Having said this, we should note that *sometimes* temporal sequence can only run in one direction, even in a passive observational study. For example, in their research on implicit egotism, Pelham and Carvallo (2015) analyzed more than 100 million 1940 U.S. Census records and showed that, for every common surname they could identify that was also an occupation (e.g., Baker, Carpenter, Farmer, Mason, Painter), men were more likely to work in an occupation if it matched their surname. There are reasonable criticisms of this study, but notice that one does *not* need to worry that men named Smith became carpenters and *then* decided to change their last names to Carpenter. In this case, there is only one likely temporal sequence. But in most *passive observational studies* (again, those in which there is no experimental manipulation), there are usually lots of tough questions about temporal sequence.

Eliminating Confounds

Even when covariation and temporal sequence are *both* obvious, researchers still can't be sure they've established causality until they address John Stuart Mill's third requirement. This third requirement is **eliminating confounds**. By this, Mill meant that we must systematically rule out all of the *competing causes* of an outcome that happen to be correlated with the cause we *think* we've identified. Do men behave more aggressively than women (a) because of evolved, biological sex differences or (b) because

of socialization? **Confounds** can be very tough to resolve. To better appreciate confounds, let's look at one that's pretty easy to debunk. Consider the finding that as the national levels of ice cream sales increase, national homicide rates also increase. Do ice cream sales cause murders? Should we outlaw the production and distribution of ice cream as a way of lowering homicide rates? Should we go a step further and actively promote ice cream spoilage? Probably not. In this case, it seems very likely that both ice cream sales and homicide rates are influenced by a third variable, namely *seasonal temperature variation*. When it's hot out, people buy more ice cream. In addition, when it's hot out, people become more easily frustrated. Frustration is known to be a strong predictor of violence, including murder. The problem of confounds is also known as the **third-variable problem**, by the way. In this case the *third* variable—besides (1) ice cream sales and (2) murder rates—is (3) heat. Heat causes changes in both ice cream sales and murder rates, and so the two variables covary with one another—and give the false appearance of a causal relation. It's hard to overstate how big a problem the third problem is if you don't take careful steps to rule it out.

Consider the ice cream example. It may look like we figured out—and logically eliminated—the confound that meant that ice cream sales masqueraded as heat. Now we know that frustration, specifically the frustration of being overheated, is the true cause of homicide. Or do we? Unfortunately, it's possible that frustration, heat, and ice cream sales are all confounded with something *else* that is the true cause of homicide. Worse yet, this true cause may be a lot less interesting than frustration. Perhaps people simply (1) drink more alcohol, (2) socialize more, or (3) get out of doors more often when it's hot out. All of these third variables (or should we say fourth, fifth, and sixth variables?) are likely to be *confounded* with temperature. Furthermore, any or all of these variables could conceivably contribute to homicides. If homicides are more likely to occur when people are drinking, socializing, or just hanging, this is a triple threat to our explanation based on frustration.

The Magic of Random Assignment

The good news is that there's a very good way to eliminate all possible confounds involving individual differences between people. The solution is to create two identical groups of research participants and study them in a true experiment. An **experiment** is a research design in which the researcher randomly assigns participants to two or more conditions, enacts a manipulation, and then assesses whether the different groups think, feel, or behave differently. The variable that is manipulated in an experiment is called the **independent variable**, and the variable that is measured (under the assumption that it is caused by the independent variable) is called the **dependent variable**. So if you manipulate frustration level in an experiment—expecting it to influence aggression—then frustration is your independent variable and aggression is your dependent variable. Likewise, if you manipulate how symmetrical human faces are, expecting people to like

the same faces more when you make them more symmetrical, then symmetry is your independent variable and liking is your dependent variable. The key to eliminating all possible confounds in a true experiment is **random assignment**. It's hard to overstate the importance of random assignment if you want to eliminate confounds. Later in this text, we'll explain exactly why random assignment does such a great job of creating two equal groups of research participants. In fact, you'll have the chance to do some random assignment yourself to see it at work.

For now, if you're wondering how something *random* can lead to a highly predicable state in which two groups of people are very similar on numerous physical, personality, and demographic dimensions, consider how predictable coin tosses are. Or if you prefer, let us show you. Take out a coin right now and toss it *exactly* 20 times, carefully tallying the exact number of heads and tails. Please write down your results (e.g., 12 heads and 8 tails). Go ahead. It should take no more than two to three minutes. Now allow us to make some predictions. Assuming you were careful and that your coin was fair, you probably tossed between 8 and 12 heads. By "probably," by the way, we mean "with a probability of .74 (74%)." Seventy-four times out of 100 when people toss a fair coin 20 times, they will observe between 8 and 12 heads. Almost 89% of the time, by the way, people will toss between 7 and 13 heads. And fully 96% of the time, people will toss between 6 and 14 heads. So we can be pretty darn sure—probably from many hundreds of miles away—that you did *not* toss 1, 2, 3, or 4, or even 5 heads. And if we wanted to be virtually positive of our predictions, we'd get you to toss the coin 100 times rather than 20 times. Our predictions would be a lot more accurate. For now, without delving too deeply into the math of random assignment, suffice it to say that random assignment works—and works *best* when you have a very large sample. In fact, in the extreme case of a tiny sample of only two people, it does no good whatsoever to flip a coin to create two "groups" of one person each. But as soon as your sample size grows to about 30 people, random assignment does a remarkably good job of creating two groups of people who are similar on almost any imaginable dimension. This is crucial, of course, because if you have created two (nearly) identical groups of people, *you have just eliminated every conceivable person confound that could exist between two different groups*. John Stuart Mill would be delighted to know that *on average* your two experimental groups were equally friendly, equally interested in politics, equally disgusted by hairballs, equally likely to have been to Disneyland as a child, equally neurotic, and equally afraid of spiders.

In the modern research era, the person who probably did the most to popularize experimentation (and thus random assignment) was a British guy named R. A. Fisher. Fisher (1925, 1935) wrote two landmark books that dramatically shaped the way social scientists think about research. As a scientist deeply interested in things like genetics and agriculture, Fisher wanted to answer questions such as what kind of manure would maximize crop yields. Plants are a lot like people by the way. Every plant is different, and this makes it difficult to know with any certainty whether a given plant grew large

because it was fertilized, because it was otherwise well tended, or because it was blessed with good genes. This meant that if Fisher wanted to study the influence of fertilizer on plant growth, he had to figure out how to create two identical groups of plants—to eliminate all possible confounds.

As you probably guessed by now, Fisher did this by popularizing the use of *random assignment*. Forgive us for being a little redundant, but using random assignment means placing specific people (or plants) in different conditions in an experiment on a totally arbitrary basis. In psychology, it means that *every participant in an experiment has the same chance as every other participant of being assigned to any condition of the experiment*. Common ways of carrying out random assignment include flipping a fair coin or using a random number generator. If you do this for a large enough group, you're virtually guaranteed to create two nearly identical groups. The best thing about random assignment is that it equalizes two or more groups on *practically every dimension imaginable*. This is the magic methodological bullet John Stuart Mill didn't know about. So if you create two groups of people—or frogs, or wart hogs—by using random assignment, you can rest assured that the two groups are identical in age, in sex, and in body mass. And if you happen to be studying people, you can rest assured that the two groups are identical on important psychological variables such as history of aggressive behavior or ice cream consumption. In short, Fisher essentially invented the experiment. In the footsteps of Fisher, experimental psychologists refined the experiment to make it a staple of basic psychological research (Aronson & Carlsmith, 1968).

So Fisher solved the problem of eliminating all possible confounds in a very clever and elegant way. He realized that two groups of people (or pea plants) can be virtually identical as groups, even though each group is made up of completely different individuals. We should add that if you are studying people rather than plants or seeds, another great way to create two identical research groups is to make each person his or own control. Within-subjects designs expose the same group of people to two or more experimental conditions—to see if people behave differently in the two or more within-in-subject treatment conditions. If they do (and if you control for important things such as the order in which people experienced the two different conditions), you'll have fulfilled all three of John Stuart Mill's conditions for establishing causality. We'll examine within-subjects designs in great detail later in this text. For now, what do traditional (between-subjects) experiments look like in practice?

EXPERIMENTS: FULFILLING MILL'S REQUIREMENTS

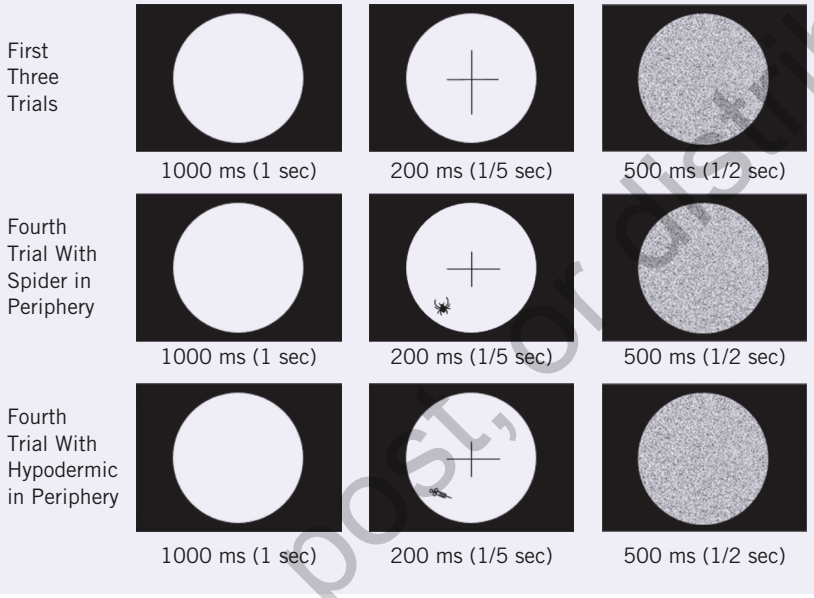
To see why John Stuart Mill probably would have loved experiments, consider a clever lab experiment by New and German (2015). New and German were interested in whether people are predisposed to detect spiders. After reviewing evidence

that venomous spiders used to be a very serious threat to human life and limb, New and German argued that if we're evolutionarily *predisposed* to stay away from spiders, spiders should be "detected, localized, and identified" more readily than other things, including other scary things that have not been around for very long in human evolutionary history. To test their hypothesis, they gave people the task of staring at the center of a circle on a computer screen. People were told that a crosshair (+) pattern would very briefly appear in the middle of the circle on each of eight trails. Participants had to press a button as quickly as possible to indicate that (a) the horizontal line was longer, (b) the vertical line was longer, or (c) the two lines were equal in length. This judgment task was more difficult than it might seem because a *mask* (a stimulus that competes with what came just before it, for short-term visual storage) replaced the crucial crosshair image after the crosshair had been on the screen for just 200 ms (that's 1/5 of a second; see the right hand column in Figure 3.2).

After making this judgment for three trials, participants repeated the task for a fourth trial. On this crucial trial, though, they were exposed not just to the circle and crosshairs but also to an unexpected stimulus—whose location in one of the four quadrants of the circle was determined at random. For *some* randomly chosen participants, the unexpected peripheral stimulus was a *spider* (see the middle row of Figure 3.2). For *other* randomly chosen participants, the unexpected stimulus was a harmless housefly (not shown). For *still others*, the unexpected stimulus was a scary but evolutionarily irrelevant hypodermic needle (bottom row). Pretest participants had reported that the hypodermic needle was just as scary as the spider. But this is presumably a fear that's learned rather than hardwired. (Ancient hominids didn't have controversial things like health care, and so they never got their inoculations.) As soon as participants made the line-length judgments for the fourth trial, the experimenter interrupted them and asked them (a) whether they had seen anything at all, other than the expected crosshairs, (b) in which of the four quadrants any unexpected stimulus had appeared, and (c) what that unexpected stimulus might have been. Participants had to choose from eight different stimuli, only one of which was correct.

Averaging across two variations on this experiment, more than half (53 percent) of those who'd been exposed to the unexpected spider were able to detect it *and* locate *and* identify it. In contrast, only 11 percent of those exposed to an unexpected hypodermic needle were able to pass all three of the same attentional tests. Results for the natural but harmless housefly were much like those for the unnatural but scary hypodermic needle. Only 10 percent of participants were able to detect, locate, and identify it. Notice that we do not have to worry, for example, that 53 percent of the participants were able to pass all three visual tests for the spider because they were much more attentive, fearful, or thoughtful than the other participants. *Variables such as attentiveness (a potential confound) should have been identical in all of the randomly assigned experimental groups.* Further, imagine that people happen to see things better than usual when things appear in the upper right hand quadrant of their visual field.

Figure 3.2 Our approximation of New and German's (2015) experimental stimuli. Note that the original unexpected stimuli were probably a little better drawn than these versions. Sometimes the unexpected stimuli were also presented closer to the center of the crosshairs, but this distance variable was held constant across different types of stimuli.



That's not a problem either because the experimenters randomly varied the location of all of the stimuli. Spiders, houseflies, and hypodermic needles all appeared *equally often* in all four quadrants. We hope you can see that experiments allow researchers a great deal of control over possible confounds. In fact, they're the only research design that controls completely for every conceivable confound involving individual differences.

Notice that in addition to eliminating all possible confounds, experiments also take care of both covariation and temporal sequence. If an experiment yields any results (e.g., 53 percent versus 11 percent detection rates), this constitutes a clear case of *covariation*. Likewise, no one has to worry about what happened *when* in an experiment; temporal sequence is always known. But now for a little bad news: It's not always possible to conduct true experiments in psychology. It's easy to manipulate whether people are exposed to spiders or needles or to make a fake interaction partner polite versus rude. But you can't randomly assign people to be male versus female or to have had nurturing versus critical parents. This problem isn't unique to psychology. Consider **cosmology** (the study of the origin and formation of the universe). Cosmologists can't do experiments to create new solar systems with different

physical properties. Instead they rely heavily on careful observations and mathematical simulations—in which they try to model what may have happened with sophisticated computer models. So you shouldn't be surprised that psychologists who study clinical disorders, human sexuality, language learning, or sensitive periods in human development often have to make clever use of nonexperimental techniques—which often use statistical rather than experimental control to rule out confounds. Now that you have a basic handle on how experiments can uncover information about causality, let's explore a few *non*experimental research designs, which often yield information that would be very difficult to gain from an experiment. In fact, most methodologists argue that nonexperimental research designs fulfill the second basic principle of making research informative. Nonexperimental research designs are often very high in *external validity*. If done well, they tell you a great deal about whether a research finding holds up in the real world.

PASSIVE OBSERVATIONAL (NONEXPERIMENTAL) RESEARCH METHODS

In addition to true experiments, psychologists make use of a very wide range of nonexperimental research designs. Although it's not possible to review them all in this summary chapter, we hope that examining just a few popular nonexperimental designs will give you a good sense of how much one can learn about psychology by using such passive observational methods. As you probably recall, the term *passive observational research design* means that researchers who use these techniques don't manipulate any variables. Instead, they have to be content to measure naturally existing variation in the variables in which they are interested. After doing this, researchers think hard about issues such as temporal sequence and eliminating confounds. So the interchangeable terms *passive observational research design* and *passive observational methods* refer to a wide variety of nonexperimental techniques for studying behavior, whether this means interviewing people, observing people unobtrusively, studying archival records, or conducting ethnographies. Let's examine a few of these passive observational methods, keeping in mind that a key strength of most passive observational methods is that they let you say a great deal about real-world behavior. In other words, these research designs have a great deal of potential to uncover information about *external validity*.

Surveys and Interviews

Surveys and interviews include research activities as diverse as conducting a national census, asking married versus unmarried romantic couples detailed questions about their sexual habits, conducting a structured interview that probes for symptoms of clinical depression, and asking people to answer three questions about a video game

on their iPhone (in exchange for some helpful clues to playing that same video game). Researchers of all stripes often try to figure out what people think, feel, and do by simply asking them. We hope you see the merits of surveys and interviews. Who knows more about you than you? You and you alone can tell us, for example, whether you ate a lot of fruit as a child, whether you physically punched anyone last week, and whether you had the fruit punch at Lara's crazy holiday party. Contrast these questions about actual day-to-day behavior with what happens in a lab experiment (e.g., staring at crosshairs on a computer screen when a silhouette of a spider unexpectedly pops up). We hope it's clear that surveys and interviews almost always focus on real-world behavior. But like experiments, surveys and interviews only yield a lot of information if researchers follow some important rules when conducting them. And it's arguably more difficult to follow the ideal rules of conducting a good survey than it is to follow the ideal rules of conducting a good experiment. This is because good surveys ideally require random sampling. Random sampling can be both technically challenging and expensive.

For example, pretty often, researchers have to be content to pose their survey or interview questions to a small group of people who happen to be handy. Such a sample is appropriately called a **convenience sample**. At the other extreme, many public opinion pollsters and at least some clinical and health psychologists often spend a great deal of time, effort, and money getting their sample just right. The ideal way to sample people in a survey or interview is to use **random sampling** (aka **random selection**). Random sampling is actually a not-too-distant cousin of random assignment. In fact, both techniques ensure that two (or more) groups of people are very similar. In the case of random sampling, though, researchers are deciding whom to study in the first place. Picking people at random from a population is the best known way to make sure that the people studied are very much like the people (perhaps millions of them) that you did *not* study. For example, a political science professor who wished to know what Americans think about climate change might *randomly* sample 1,000 Americans to see what they think. If the researchers carefully identified their target population (e.g., registered voters or current adult residents of the United States) and randomly selected 1,000 such people, the researchers could be pretty confident that the 250 million American adults they did *not* have time to sample were pretty similar to the 1,000 adults they did sample.

Later in this text, you will learn a great deal about why random sampling is so important. The gist of this upcoming conversation is that if you only study convenience samples or if many people refuse to take part in your study in the first place, you'll probably end up with a sample that does not really resemble your population of interest. For example, if you only call people up for a phone interview using land lines (remember those?), your sample will probably skew much older than the current U.S. adult population. As another example, if you only sample college students in a classroom, you will have introduced the opposite bias (a very young sample).

These are both examples of **selection bias** (also referred to as *sampling bias*). This means sampling people in such a way that your participants do not represent well the population of people whose opinions you were hoping to measure. It's practically impossible to avoid all possible forms of sampling bias. Having said this, any sample is better than no sample at all. Further, some imperfect samples are better than others. If your college or university has a very diverse student population, a convenience sample at your school is a little more impressive than a convenience sample of, say, college students at a school whose enrollment is 95 percent White. Even under less-than-ideal conditions, however, surveys and interviews can still be incredibly useful. For example, if your survey of college students shows that sociology majors say they are more likely to vote Democratic than economics majors, this will tell you very little about whether a Democrat or a Republican is more likely to win the 2032 U.S. presidential election. However, it may still give you some nice insights into how students interested in sociology differ from students interested in economics. Don't forego the chance to do a survey or interview because you're stuck with a convenience sample. But if you are stuck with a convenience sample, be ready to think hard about and acknowledge this methodological limitation.

In addition to having to grapple with sampling, those who conduct surveys or interviews face at least two other major methodological problems. These problems are (a) that people are not always *able* to report their experiences honestly (e.g., because of fallible memories or language barriers) and (b) that people are not always *willing* to report their experiences honestly (e.g., because of social desirability biases or legal worries). If you don't believe us, then send your mom a text message telling her exactly how many times you smoked pot last week. Luckily, there are some pretty good ways to minimize these two problems. One way is to be sensitive to the issue of time. People can better remember exactly what they did yesterday than exactly what they did as children. People also respond more honestly when they know that their answers will be kept completely **confidential** or private (Schroder, Carey, & Vanable, 2003). One very clever way to increase truthful responding to surveys—in the lab, at least—is known as the **bogus pipeline**. The inventors of this technique, Jones and Sigall (1971), convinced participants that they had invented a highly accurate "lie detector." This deception worked, by the way, because—unbeknown to the participants—Jones and Sigall already had access to the participants' true attitudes on several issues. Thus, when Jones and Sigall were presumably just "calibrating" the machine, participants observed what seemed to be striking evidence for the machine's accuracy at lie detection. When White men believed that researchers could truly read their minds, they reported attitudes about Black Americans that were significantly more negative than those reported by White men who just responded to a traditional written survey. Incidentally, attention to detail is important when using the bogus pipeline. As Roese and Jamieson (1993) showed in their careful review, fake lie detectors work best when you ask people *what the machine will say* about their attitudes—*not* when you simply ask people to report what their true attitudes

are while they're hooked up to the machine. Ironically, then, one of the best ways to get people to tell the truth about sensitive subjects is to lie to people.

Of course, bogus pipelines are a very real pain to use. You can't use them, for example, if you haven't pretested people so that you know their attitudes about a wide range of topics. Are there any easier ways to increase honest responding in surveys? There are, and one of these ways becomes especially important when you're assessing sensitive topics such as sex, aggression, clinical disorders, stereotyping, or even self-esteem. Schroder and colleagues (2003) found that the specific mode of conducting a survey can matter a lot. People appear to give much more honest answers when they fill out self-administered questionnaires than when they do face-to-face interviews. Arguably, the growing trend toward collecting survey data on the web—to the extent that it relies on self-administered questionnaires—could be a very positive development. We should quickly add that this does *not* mean one should diagnose depression or schizophrenia using self-administered questionnaires rather than clinical interviews! One of the reasons people spend several years in graduate school to get an advanced degree in clinical psychology is to master clinical interview techniques and extract information that would be difficult, if not impossible, to extract from a written survey. But if you're doing a simple survey about political attitudes or sexual preferences, you'll probably get more honest answers if you let people fill out written surveys privately rather than interviewing them face-to-face. Finally, for traditional research on attitudes or the self-concept, things as simple as putting a mirror in the room where people are filling out their surveys can also increase accurate responding (Duval & Wicklund, 1972). A full review of all of the ways to avoid bias in survey responses is beyond the scope of this chapter. Suffice it to say, though, that there are some pretty good solutions to most problems with surveys and interviews. Researchers who follow a few simple rules about conducting good surveys and interviews can gain a lot of information about what people think, feel, and do in their daily lives.

Unobtrusive Observation

When a behavior is sensitive enough that many people won't readily admit the truth about it or when people's memories prove to be poor, the best way to figure out what people do may be to engage in **unobtrusive observation**—that is, *to record their behavior when they don't know you're doing so.* This can be ethically tricky, of course, but if a person is in a public place with no expectation of privacy, most ethicists would say it's OK to observe the person, especially if the observer makes sure no harm or embarrassment comes to the person being observed. The two main keys to making good unobtrusive observations both have to do with *keeping them* unobtrusive (i.e., secret). Observations are truly unobtrusive only if (a) researchers themselves don't interfere in any way with people's natural behavior and (b) research participants don't have any idea they're being observed.

A great example of unobtrusive observational research is **garbology**. Rather than surveying people about what they buy, some marketing researchers retrieve people's trash (and/or recycling) before the pickup crews can get to it. So if you're worried that people will over-report their kale consumption or underreport their *ale* consumption, you can spend a lot of time sorting through people's trash and carefully count up the number of kale bags and ale bottles that turn up. In the age of Google, it is also possible to let others do some of the counting for you. If you want to know what Americans search for in Google, for example, you can check out the free research tool known as **Google Correlate**. The helpful folks at Google keep very careful tabs on Google search volume for millions of search terms—both over time and across U.S. states. As you might guess, Americans search for terms such as "common cold" and "ski trip" much more often in the winter months than in the summer months. People also search *much* more often for "sunscreen" in Hawaii than in Alaska. If this sounds like an exercise in the obvious, consider a recent study by Pelham and colleagues (2018). Inspired by *terror management theory* (which suggests that people spend a lot of time and energy managing their fears of death), this research team analyzed week by week changes in Google search volume for (a) terms for life-threatening illnesses ("cancer," "hypertension," and "diabetes") and (b) terms reflecting an interest in religion (e.g., "God," "Jesus," and "prayer"). Even after controlling for seasonal, annual, and holiday-based variation in this indicator of the public's online interest in religion, Pelham and colleagues found that when searches for major illnesses increased in one week, searches for religious content increased the next.

As we were writing this chapter in December of 2017, a great example of unobtrusive observational research was getting a great deal of media attention. Audrey Blewer et al. (2017) analyzed data from more than 19,000 cases of cardiac arrest in the United States. Unfortunately, Blewer found that men were significantly more likely than women to receive cardiopulmonary resuscitation (CPR) from a bystander when they were unlucky enough to experience cardiac arrest in public. Partly for this reason, men also had higher survival rates than women did in this unhappy situation. Blewer speculated that in a public setting in which the victim is a stranger, many people may be highly reluctant to touch a woman's chest—even if the reason for so doing is to save the woman's life. One piece of evidence that suggested Blewer may be right is the finding that when people went into cardiac arrest around people they *knew*, the gender difference in receiving CPR evaporated. We would hazard the prediction that if you simply asked people if a fear of touching a woman's chest would keep them from saving her life, very few people would answer in the affirmative.

If you want to make a less sensitive unobtrusive observation of your own over the next few weeks, just pay careful attention to a few gasoline pumps. In particular, take a close look at the buttons people press to select octane levels. Unless the stations where you shop for gas have brand new pumps, you'll probably see that the button for one of the three grades (87, 88–90, or 91–94) has a lot more wear and tear than the other two buttons. What does this tell you about consumer price preferences? Would

this unobtrusive observational finding vary in rich versus poor neighborhoods? Another unobtrusive observation you could make with a bit more forethought would be to see how often a person holds a door open for the person entering a building behind them (e.g., "At the main entrance to Gravenor Hall, between the hours of . . . , people held the door open 83 percent of the time when another person was 3 meters or less behind them."). Notice that observing live human behavior would be more precise than estimating wear and tear on buttons, but you'd also have to take some precautions to be sure people didn't know you were observing them. Once you solved these problems, you could get a lot of data. For example, it'd be easy to code for the age, gender, or apparent ethnicity of the person holding the door—and for the person who would benefit from having the door held. The list of variables for which you could code is limited only by your dedication and imagination. Is it raining out? Are classes just about to start, or is there still 10 minutes to go? Is the person in need of door holding physically attractive? Physically disabled? Sometimes the best way to know what people really do is to make careful, unobtrusive observations. Notice that in all of the cases mentioned here, we're measuring real behavior in the real world. We may not have great control over confounds or even temporal sequence, but in most cases of unobtrusive behavioral research, we know that the behavior in question is real.

Archival Research

A pretty close cousin of unobtrusive observational research is **archival research**. This is *research that uses existing public records to test research hypotheses.* The social cognitive study of men named Baker or Carpenter qualifies as archival research, as did the study using Google Correlate data. The big difference between archival research and unobtrusive observational research is that, in the case of archival research, someone else has already done a lot of the observing for you. All you need to do if you are an archival researcher is to gain access to the public records that someone else has gathered. One of the best examples of archival research in the social sciences is also one of the most chilling. Martin Daly and Margo Wilson (1994, 1998) have conducted a great deal of archival research on murder, including research on who most often kills children. A very distressing fact about child homicide is that parents who do not share any genes with the children for whom they care (stepparents) are much more likely to kill children than are parents who do happen to share genes with the children (biological parents). On a happier note, a great deal of archival research by Manuel Eisner (2003) has shown that over the past several centuries, we human beings have become much less likely to murder one another. This radical reduction in murder rates over time seems to be the result of many different processes, including the development of economic and political systems that allow people to profit without taking the land or possessions of others (see Pinker, 2010). But evolutionary psychologists such as Steven Pinker have been quick to note that it wouldn't be so easy to civilize people if

we weren't inherently predisposed to civility (at least under the right circumstances). One nice consequence of the rapid development of the internet over the past couple of decades is that an amazing amount of archival data are now available to the public. And a great deal of these data are free. Later in this chapter, we'll introduce you to a couple of additional examples of archival research.

Ethnographies

Archival research is great—when you can gain easy access to reliable archival data. But one can only do archival research on topics that have been archived. There are plenty of records about who marries whom and who kills whom. But no one keeps public records of exactly how people reason about conflict, how close people stand to one another as they speak, or how respectful people are of their elders. When researchers want an in-depth look at how people think, feel, and behave, especially in cultures about which researchers know very little, they often conduct ethnographies. As Bernard put it, an **ethnography** is "a narrative that describes a culture or a part of a culture" (2006, p. 34). But as Bernard would be quick to add, there are *many* kinds of ethnographies—ranging from those in which the ethnographer is very careful not to influence those he or she is observing at all (think Jane Goodall, who initially did not interact at all with the chimps she observed) to *participant ethnographies*, in which the ethnographers truly embed themselves in the culture they are studying—under the assumption that experiencing something yourself is the best way to understand it well.

In some ways, good ethnographies are the opposite of good archival research. For example, ethnographies often require a lot of behavioral coding on the part of the ethnographers. In archival research, though, someone else has always done the coding for you, like it or not. Because ethnographers are in the physical presence of those they observe, they may also have to go to great lengths not to change what they are observing. Otherwise, all they have is a good account of what people do when they know a stranger is watching them. This is not an issue at all in archival research. Ethnographers may also have to spend months, if not years, to learn the language and customs of those they study. In contrast, archival researchers just need to be sure they have gotten hold of the right records. As a final example, archival research often focuses on very large groups of people, such as entire states or nations. In contrast, by necessity, most ethnographies focus on one small group, whether it is a small tribe of people living in Papua New Guinea or a small group of U.S. consumers (Mariampolski, 2006). Archival research often tells you just one thing about a huge and familiar group of people. Ethnography often tells you many different things about a tiny and unfamiliar group of people. What the two techniques have in common (besides being passive observational methods) is that they both yield information you could never gain in the lab.

TRADE-OFFS BETWEEN INTERNAL AND EXTERNAL VALIDITY

So far we hope it's clear that experiments provide a lot of information about internal validity. We hope it's equally clear that passive observational research designs often provide a lot of information about external validity. But so far, we have emphasized only the wonderful features of both experiments and passive observational studies. Each of these two kinds of research also has its major drawbacks. The same experiments that are usually very high in internal validity can often be very low in external validity. Likewise, the same passive observational designs that are usually high in external validity can often be very low in internal validity. It shouldn't surprise you that there are trade-offs in research. Making a tool better at doing one job often makes it worse at doing another. Crescent wrenches make terrible screwdrivers. Formula race cars get you where you're going quickly. But the trade-off in fuel and safety isn't worth it to most commuters. Life is full of trade-offs. What might surprise you, though, is that the well-known trade-off between experimental and no-experimental research—that is, the trade-off between internal and external validity—is not always *quite* as bad as it seems at first blush. Careful researchers can sometimes design lab studies that actually do quite well on the dimension of external validity. Passive observational researchers can sometimes design archival studies or surveys that do a surprisingly good job of addressing internal validity. To explore these ideas, let's examine exactly what methodologists usually mean by internal and external validity. The better we understand these terms, the better a job we can do of trying to maximize them—in any kind of research design. We'll begin by tackling the most difficult of John Stuart Mill's three requirements for establishing causality, and we'll argue that all confounds are not created equal. Then we'll deconstruct external validity. We'll argue that external validity usually boils down to one of four separate concerns about the generalizability of a research findings. If you know what the four possible concerns are, you're in a good position to do everything you can to maximize external validity. Moreover, this is true regardless of whether you are an experimenter, an archival researcher, or a survey researcher. Let's begin by taking a close look at confounds.

CAGES: THE "BIG FIVE" OF WORRISOME CONFOUNDS

As you now know, well-conducted experiments eliminate a virtually infinite list of possible confounds. This implies that the list of possible confounds that threaten *non*-experimental research is endless. We are happy to say that it's not quite that bad. It looks like there are just five worrisome confounds that show up over and over again

in a great deal of nonexperimental research. This is very good news. After all, if some confounds prove to be much more common than others, researchers who focus on the most common ones will often go a long way toward solving the third variable problem. Along these lines, there is good reason why census takers, anthropologists, epidemiologists, sociologists, clinical psychologists, and marketers have long focused on a handful of regional and demographic variables when doing their jobs. Five of the cardinal ways in which human beings vary include geography, age, gender, ethnicity, and socioeconomic standing (SES, education and/or income). Following Pelham (in press), we refer to these five key variables (all of them potential confounds) using the acronym **GAGES**. So if you could eliminate just five confounds in a new research program, we suggest that eliminating concerns about geography, age, gender, ethnicity, and SES would be a great start. Let's take a quick look at each of these five variables, examine why they are all so worrisome, and thus see why it is so great if we can address all of these concerns in nonexperimental research.

Geography

Geographically speaking, knowing where a person lives can be very telling. From red states versus blue states to latitude versus altitude, location matters. According to the 2010 U.S. Census, the average Maryland resident had almost twice the family income of the average West Virginia resident. As another example, New Jersey is about 1,000 (yes, 1,000) times more densely populated than Alaska. Personal beliefs and values also vary widely across U.S. states. Residents of Vermont are more than five times as likely as residents of Mississippi to report that they are not religious (Newport, 2014). People who grow up in Georgia or Alabama are much more likely to be sensitive to threats to their family honor than people who grow up in Michigan or New York (Nisbett & Cohen, 1996). In fact, research on cultural evolution suggests that properties of the physical environment predict variables as different as what kind of language people speak, whether people cook with spices, whether women are allowed to have multiple husbands, how much parents value obedience, and xenophobia (Billing & Sherman, 1998; Everett, Blasib, & Roberts, 2015; Murray & Schaller, 2014). Some of these geographic realties are obvious. People who live in Hawaii get much more sun than people who live in Alaska. But other geographic confounds aren't so obvious. For example, people living in colder U.S. states are typically more skeptical of the reality of global warming than people living in warmer states (Pelham, 2018c).

Age

Demography can matter just as much as geography. Beginning with age, older Americans worry less than their younger counterparts (Newport & Pelham, 2009).

They also eat healthier diets, exercise less frequently, and care more deeply than young people do about nurturing close, established relationships (Carstensen, Isaacowitz, & Charles, 1999; Dugan, 2013). Older Americans are also substantially more likely than their younger counterparts to be religious and to be wholly unfamiliar with Lil Wayne. Adolescents are also higher in both risk taking and egocentrism than are young adults (Blakemore, 2012). A person's risk of both suicide and cancer also varies dramatically across the lifespan. Differences such as these are why there is an enormous field of research called lifespan developmental psychology.

Gender

Moving on to gender, across the globe, men are more likely than women to assault or kill others, to commit suicide, to work in dangerous jobs, and to abuse drugs. Conversely, women are more likely than men to suffer from depression and to serve as caretakers, both at home and at work (e.g., as nurses). On average, women also earn less money than men. On average, men and women may also think about religious and moral issues pretty differently (e.g., see Miller & Hoffman, 1995; Winseman, 2002). The list of ways in which gender matters is so long that there is an entire branch of research in the social sciences known as gender studies. It is also important to note that, like the other GAGES variables, gender influences not only how we behave but how we are routinely treated by others. Remember the study of whether people suffering from cardiac arrest received much-needed CPR? Gender should not have mattered. But it did. Remember Donald Trump and Harvey Weinstein? Many have argued that they treat men and women very, very differently.

Ethnicity

Ethnicity matters, too. Both Blacks and Latinos are more likely than Whites to suffer from clinical depression (Dunlop, Song, Lyons, Manheim, & Chang, 2003). Relative to Whites, Blacks are also much more likely to lack confidence in the police (Jones, 2015), more likely to vote Democratic, and much, much more likely to be familiar with Lil Wayne. More than 50 years after the passage of the U.S. Civil Rights Act, there are still large ethnic differences in income, unemployment, incarceration, and education. On a brighter note, especially if you grew up adoring your *tatara abuela*, Latinos in the United States live a bit longer than Whites do, despite being substantially more likely to live in poverty. Many ethnic minorities are also substantially more likely than Whites to be immigrants and to be bilingual. If you are studying language use or vocabulary, for example, ethnicity is a potential confound you would want to take very seriously.

Socioeconomic Standing

Above and beyond ethnicity, one of the best predictors of longevity and well-being is SES. Loosely speaking SES means wealth plus education (Bosworth, Burtless, & Zhang, 2015). SES also predicts important attitudes and values (Pelham, in press) as well as serious problems such as suicide risk and automobile accident rates (Sehat, Naieni, Asadi-Lari, Foroushani, & Malek-Afzali, 2012). One distressing aspect of SES is that, when people live in poverty for a long time, they sometimes come to internalize the idea that their work is not very valuable (Pelham & Hetts, 2001).

Research on life history theory also shows that growing up poor (regardless of how rich you are as an adult) makes people more likely to accept the philosophy that "Life is uncertain. Eat dessert first." Growing up in a world of financial uncertainty is probably one reason why people who grow up poor get married at a younger age and have more children than people who grow up wealthy. There is a reason why economists, sociologists, and political scientists all study SES. It really matters.

Given the importance of the GAGES, researchers who conduct nonexperimental research will ideally be able to show that their research finding goes above and beyond any of the confounds summarized by GAGES. Later in this chapter, we'll delve into this idea in more detail. For now, let's consider a single example. More than 30 years ago, when the first author of this text was taking a first course in research methods, Ed Vatza introduced him to a research study that had recently gotten a lot of attention—especially among music lovers. The study showed that orchestra conductors live a *lot* longer than people in other occupations. There was a lot of speculation and debate about exactly why this is the case. Maybe there is something healthy about the *mental stimulation* of thinking about music all day. Maybe *emotional* aspects of exposure to music are good for you. Maybe orchestra conductors tend to be *beloved*, and maybe all that love is good for your immune system. Maybe. But now that you know about GAGES, can you see how being an orchestra conductor is *confounded* with some very important GAGES variables that are known to influence longevity? Consider geography. Orchestra conductors tend to live in the city. If city dwellers live longer than rural dwellers, this is a big confound. It's also the case that most orchestra conductors are male. But if one did not control for gender, notice that being male is a liability, *not* an advantage, where longevity is concerned. So in this study being male is actually a **reverse confound** (a confound that makes it *harder* than it would be otherwise to observe an effect rather than masquerading as the effect). Orchestra conductors tend to be White and wealthy, and either of these GAGES variables is a big potential confound. But the confound that takes the cake is age. Further, the cake in question is probably a *birthday* cake because people only become orchestra conductors in the first place when they are pretty old. There is certainly nothing that is more badly confounded with longevity than age! From this perspective one "occupation" in which people probably live even longer than orchestra conductors is "great grandmother."

What is it about great grandmotherhood that makes them live so long? Maybe it's all the love they receive, or maybe it's the age requirements of being one in the first place! In any passive observational study, this is the kind of careful analysis one must conduct before concluding that an independent variable (like being an orchestra conductor) has a true effect on a dependent variable (like longevity). GAGES is a simple rule of thumb for stimulating this kind of critical analysis.

Of course, the list of possible confounds about which researchers should worry does *not* end with GAGES. Specific confounds vary with the specific research question at hand. In research on health psychology, a passive observational researcher might need to control for smoking rates, exercise levels, diet, chronic stress, and social support levels. In research on language learning, one might need to consider whether a toddler's mother is skilled at "motherese" (e.g., speaking in ways that draw attention to novel words). But the five major worries summarized by GAGES are a great place to start in any nonexperimental research design. Perhaps the most important implication of the GAGES heuristic for nonexperimental research is that *even if time and survey space is very short, one should always measure the GAGES variables.* At a minimum, this will make it possible to see exactly how problematic the GAGES variables prove to be in a specific research study. In short, then, if you are familiar with some of the most common confounds that crop up in psychological research, you are in a good position to try to address them and bolster the internal validity of a passive observational study.

EXTERNAL VALIDITY AND THE OOPS! HEURISTIC

Just as it is helpful to have rules of thumb for analyzing internal validity, it is useful to have rules of thumb for analyzing external validity. The concerns methodologists raise when they question the external validity of almost any social scientific research finding seem to fall into only four categories. The concerns include questions about operational definitions (**operationalizations**), questions about generalization based on time (**occasions**), questions about generalization with respects to **populations**, and questions about generalization of an effect in different **situations**. With this in mind, Pelham (in press) suggested the **OOPS! heuristic** as a convenient way to summarize these four concerns about external validity. What, exactly, are these four specific concerns?

Operationalizations

We can only study things scientifically if we devise clear and precise operational definitions of those things. But there are many different ways to operationalize most hypothetical constructs. In men at least, one operational definition for sexual arousal

is based on volumetric changes in the penis. Another is based on genital temperature, based on blood flow. A third is based on simple self-report. Notice that an advantage of the second approach (based on genital temperature) is that it should work about equally well for women and men (Kukkonen, Binik, Amsel, & Carrier, 2007). When multiple operational definitions of something are all reasonable, we can place greater confidence in a finding when the finding holds up well across *all* of these different operational definitions. *Altruism*, for example, could be defined in terms of either (a) giving food or physical resources to a fellow organism or (b) risking your own safety to protect a fellow organism from harm. Nursing fits the first definition of altruism. Making an alarm call ("There's a hawk up there!!!") fits the second. Even something as basic as romantic attraction can be operationalized many different ways. How close you sit to another person, how much time you spend looking into his or her eyes, how much you say you would like to kiss the person, and whether you are married to the person (versus unattached or divorced) could all be considered reasonable indicators of romantic attraction.

Likewise, as you may recall from Chapter 1, we could operationalize hunger by simply asking people how long it has been since they last ate. Alternately, we could ask people to rate their current level of hunger on a 9-point scale, where 1 is "not at all hungry" and 9 is "extremely so." An advantage of selecting hours of food deprivation is that this particular operational definition works for human adults, human infants, and naked mole rats. Whatever topic or species you are studying, evidence for anything is more impressive when this evidence holds up across many different operational definitions. One reason why Steven Pinker's (2010) argument that human violence has declined over the past few centuries is so convincing is the fact that Pinker uses *many*, *many* different operational definitions of violence, from killing people or cutting off their noses to enslaving or imprisoning people. He also examines data on warfare, child abuse, burning witches, and hurting animals. Across a vast range of operational definitions, violence has dramatically declined.

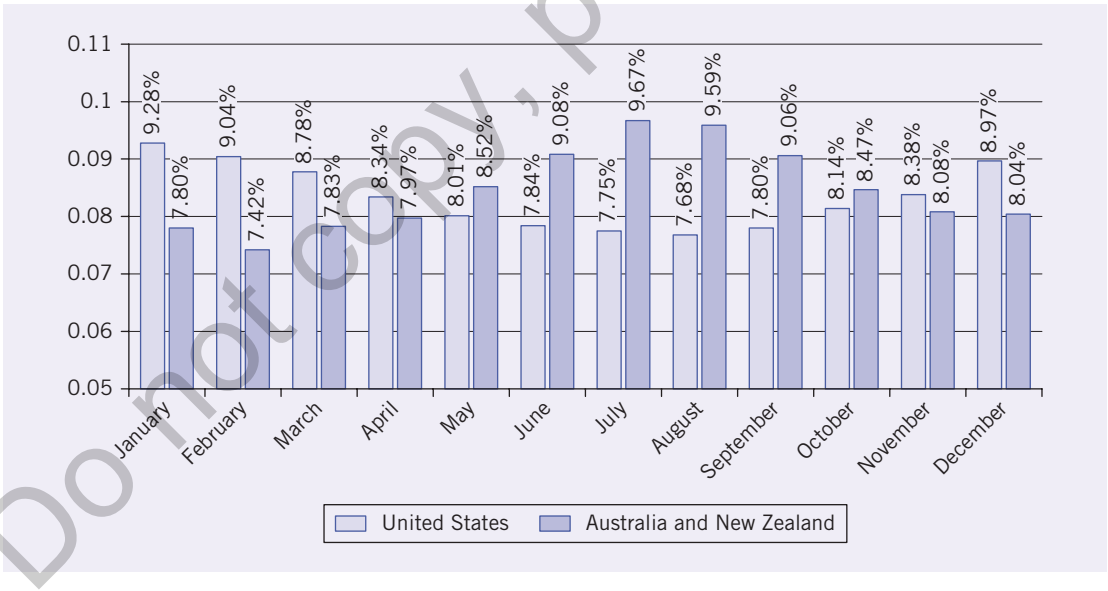
Occasions

"To everything there is a season." Indeed, human behavior has always varied greatly across the day-night cycle, across the seasons, and across the millennia. Roberto Refinetti (2005) found that college students were much more likely to report having sex late at night than at any other time of day. About half of all the sexual interactions his participants reported in a three-week daily-diary study took place during the two-hour window between about 11 p.m. and 1 a.m. Many other things, including people's hormone levels, vary naturally over time. And this natural variation often has important consequences. For example, Lisa Welling and colleagues (2008) showed that men's rated attractiveness of highly feminine as opposed to less feminine female faces was stronger than usual on days when the men's testosterone levels were higher than usual.

Many more world records in track and field are set in the early evening than in the early morning. Academic performance also varies with the time of day. As children move from middle childhood to adolescence, they tend to stay up later at night and have more difficulty waking up early. In fact, research shows that one of the easiest things U.S. educators could do to improve academic performance among high schoolers would be to start high school an hour or two later. Experiments in which educators have tried this have shown substantial academic gains among high schoolers who get to start school later in the morning (Minges & Redeker, 2016). So unless you are truly a "morning person," you'd be better off scheduling that dreaded Calculus III class for 2 p.m. rather than 8 a.m.

Looking at timing over a broader window, both human births and human deaths vary with the seasons. The archival data in Figure 3.3 show that Americans more often die in winter than in summer (despite the fact that deaths by accident are more common in the summer; Rozar, 2012). There is debate about exactly why this seasonal pattern occurs, but the pattern is clearly *seasonal* rather than calendrical. The pattern reverses in the Southern hemisphere. Marriage rates, too, vary over the course of the year. As you probably knew, June is the most popular month for U.S.

Figure 3.3 Likelihood of dying by month in the United States versus Australia and New Zealand, 1990–2010, adjusted for the number of days in a month. U.S. Data source: Social Security Death Index (SSDI). Australia/New Zealand data source: "Find a Grave Index" (sources accessed at ancestry.com).



Source: U.S. Data from Social Security Death Index (SSDI); Australia/New Zealand data from "Find a Grave Index," accessed at Ancestry.com.

weddings. Late summer months are also pretty popular. Perhaps you could have guessed, then, that the deep winter months are the *least* likely months for weddings. As you almost certainly did *not* know, people are more likely to get married during the month of their own birthdays than in other months (Pelham & Carvallo, 2015). And yes, months of birth also vary with the season. More U.S. babies are born in September than in any other month.

Time also matters century by century. Two thousand years ago, Romans died more often in the summer than in the winter (because diseases like malaria were much more common in summer; Scheidel, 2009). At that time, the entire population of the earth was smaller than the current population of Indonesia. Turn the clock back to 10,000 years ago, when agriculture barely existed, and the earth's human population was smaller than the current population of *Chicago.* So time matters. For this reason, when evaluating any research finding, we have to ask ourselves if that finding would hold true at other times. Showing that something interesting is true is impressive. Showing that it was *also* true 150 (or 150,000) years ago is even more impressive.

Populations

Almost no research finding, psychological or otherwise, applies to every imaginable population. People see colors much more vividly than dogs do. Dogs smell things people can barely imagine. Even if we limit ourselves to people, the external validity of a specific research finding can vary dramatically depending on the population. Presumably very few devout nuns or Buddhist monks believe it is appropriate to kick a person's butt if the person insulted their mothers. But many Southern men certainly feel that way (Nisbett & Cohen, 1996).

So research findings become more impressive when we learn that they hold up across many different populations. Notice we said "hold up across many different populations." Simply having a diverse population is nice, but it's a lot more impressive to show that an effect holds up across all of the interesting subpopulations you have studied. This will often require a very large sample size. Consider an example where very large sample sizes *are* often available. Research on age preferences in marriage has shown that in almost every culture ever studied there is an average tendency for women to marry men who are at least a year or two older than they are (Demetriou & Pollet, 2015). We also know of no cultures in which women are more physically violent than men. Some researchers even test their hypotheses in multiple species rather than multiple human populations. Long ago, Bob Zajonc (1965) argued that organisms tend to run, jump, eat, or fight harder when they are in the presence of other members of their species than when they are alone. Zajonc (rhymes with "science") dubbed this phenomenon *social facilitation.* And Zajonc (rhymes with "alliance") tested and confirmed his hypotheses about social facilitation in ants, cockroaches, parakeets, puppies, and monkeys, as well as in people. That's a pretty diverse overall population.

Situations

A final aspect of external validity has to do with generalization across different situations. All research takes place in a specific context, and that context may dramatically influence what researchers observe. The way people think and reason seems to vary based on the way in which an experimenter dresses. When experimenters talk and dress casually, people seem to *think* casually (i.e., more intuitively, less logically; see Simon, Greenberg, Harmon-Jones, Solomon, Pyszczynski, Arndt, & Abend, 1997). And when people dress formally, others are much more likely to obey them (Bickman, 1974). Situations can be hard to separate from occasions (in a sense, time of day is a situation), but situations are not the same as occasions because situations can vary separately from time (e.g., experimenters can be formally or informally dressed in both January and July; participants can be placed under time pressure both in the morning and in the evening). In short, then, to know how robust a research finding is, you need to know how well it holds up in a wide variety of situations.

Are GAGES and OOPS! WEIRD?

If you have ever had a course in cross-cultural psychology, you may be wondering if GAGES or OOPS! is the same as the **WEIRD** critique. Henrich, Heine, and Norenzayan (2010) argued that a great deal of psychological research fails to consider the tremendous cultural diversity of the planet. Specifically they noted that the great majority of past research in psychology focused on WEIRD people, namely those who come from "Western, educated, industrialized, rich, and democratic" societies. But GAGES is distinct from WEIRD. First, WEIRD expresses a concern about maximizing *external* validity (e.g., would shopkeepers in India behave like students in Indiana?). By contrast, GAGES is all about maximizing *internal* validity. That being said, GAGES does have some overlap with WEIRD. Cultures, after all, have geographies. Cultures also vary in age, ethnicity, SES, and even gender ratios. Furthermore, one could easily treat GAGES variables as cultural moderators rather than confounds. Conceptually, however, WEIRD overlaps more with OOPS than with GAGES. One key difference here is that WEIRD focuses on *cultures* whereas OOPS! usually focuses on *individual people*. Further, WEIRD includes populations and situations but is largely silent regarding operationalizations and occasions. In a sense, then, OOPS! means that the WEIRD critique is extremely useful but may not go quite far enough.

Getting back to the OOPS! heuristic, to the degree that a specific study or empirical report shows that an effect holds up well using different operational definitions, in multiple temporal windows, for different populations, and in different situations, there can be little doubt that the study or report has a great deal of external validity. Is there a good way to maximize external validity while also minimizing threats to internal validity? Although this is a very high methodological bar, we'd like to argue that the answer is yes, for two reasons. First, you can conduct true experiments or approximations

thereof in which you work hard to consider the four basic issues summarized in the OOPS! heuristic. This might mean studying two or more very different *populations*, or capitalizing on a single, very diverse population. It might mean making use of more than one *operational* definition of your independent and/or dependent variables. You might also consider varying the *occasions* on which or the *situations* in which you conduct your experiment. In short, you might strive to maximize the external validity of a set of experimental research findings. Second, you can conduct nonexperimental research and work very hard to address confounds such as those summarized by the GAGES heuristic. In fact, even researchers who conduct the most passive of all forms of passive observational research—archival research—can often do a very good job of addressing the concerns summarized by GAGES. In the next section of this chapter, we'll consider a classic experimental research finding in social cognition and take note of how the person who documented this finding worked hard to maximize *external* validity. Following this example, we'll offer three examples of archival research in social cognition that all do a nice job of minimizing the GAGES confounds and thus maximizing *internal* validity.

OOPS! HE DID IT AGAIN: MAXIMIZING EXTERNAL VALIDITY IN THE LAB

Mere Exposure

One of the most influential psychologists of the 20th century was Bob Zajonc (the same Zajonc who did lots of externally valid laboratory research on social facilitation). Further, one of the findings that helped solidify Zajonc's reputation as a scientific genius (besides the fact that his last name is practically "s*cience*") was the **mere exposure effect**. The mere exposure effect (Zajonc, 1968, 2001) *is the finding that the more often people are exposed to something, the more they usually like it*. This idea was pretty controversial when Zajonc proposed it in the 1960s because, back then, learning theories that emphasized reinforcement and punishment were the dominant viewpoint in psychology. In fact, some have argued that the mere exposure effect also flew in the face of the emerging cognitive perspective that was challenging some of the cherished principles of learning theories. So Zajonc had his work cut out for him when he proposed the hypothesis that *merely exposing* people to something (without rewarding them at all) would usually make them like it. Zajonc also had his work cut out for him trying to convince readers that people didn't need to be *aware* of having been exposed to something a lot to like it a lot. Why the heck should people like something just because they had seen, felt, heard, tasted, or smelled it before? And there was a third problem with Zajonc's hypothesis, namely *reverse causality* (aka temporal sequence). It was already very well known and very obvious that people choose to expose themselves to things they *like* a lot more than people choose to expose themselves to things they

dislike. People don't like specific foods because they've eaten them; they choose to eat specific foods because they like them. At least that's what everyone assumed before Zajonc did his classic work on the mere exposure effect.

Again, as the name of the effect implies, Zajonc hypothesized that *merely exposing* people to something—in the absence of any reward—makes people like it. The reason Zajonc was able to convince skeptical readers that the mere exposure effect is real—and that it influences a wide range of real world judgments—is that Zajonc had a great appreciation of both internal and external validity. Zajonc did *not* merely show, for example, that people like words and letters much more than usual when the words or letters in question are common rather than rare in a language. Zajonc took care of internal validity by conducting experiments to document the mere exposure effect. In one of his early experiments, Zajonc (1968) pretended to be studying language learning. He asked college students to try to pronounce what they thought were Turkish adjectives. The number of times people were exposed to the fake Turkish adjectives (e.g., "zabulon") varied from 1 to 25 exposures. In addition to the fact that all the fake words were seven letters long and all had three syllables, Zajonc made sure that each of the 12 fake words he studied appeared with equal frequency to the total sample of participants. (This is known as *counterbalancing*, and you'll learn the details of this clever technique later in this text.) This meant, for example, that Luke might see "lokanta" only once but see "kadirga" 25 times. Meanwhile, Sara might see "kadirga" only once while seeing "lokanta" 25 times. In this way, Zajonc could separate any effects of the fake words themselves from the true effects of exposure. *For all 12 of the fake Turkish words, people liked the words more when they had been exposed to them more often.*

If this experiment leaves you with any concerns about external validity, please hold them momentarily. Zajonc was just getting started. Let's start with operationalizations. Zajonc and others have documented that mere exposure applies to geometric shapes and fake Chinese ideograms as well as fake words. They also showed that some of their nonexperimental effects apply to real German, French, and Spanish words as well as to types of words as varied as bird names, tree names, city names, and personality trait terms. Zajonc and others also showed that mere exposure effects apply to liking for musical tones, classical music pieces, yearbook photographs, paintings, novel color combinations, and food preferences. Further, the effect appears to get larger with greater numbers of exposure. All else being equal, people like something they've seen 10 times more than something they've seen five times—and something they've seen 100 times more than something they've seen 10 times.

On the OOPS! dimension of *occasions*, there is no strong reason to think mere exposure effects would wax and wane with time of day or month of the year. But researchers have shown that the effect travels well across time—and that the effects of mere exposure are long lasting. Some of the informal demonstrations of mere exposure involved data sets that were collected and tabulated long before Zajonc published his initial reports. Further, the mere exposure effect has now been replicated over a

window of more than 50 years since Zajonc's first studies. In fact, some of the most impressive evidence for mere exposure comes from studies that show that greater exposure to stimuli in the womb (or prior to hatching if you are a bird) increases preferences much later in an organism's life. Babies whose mothers read Dr. Seuss's *The Cat in the Hat* to them twice a day *in utero* (before birth) for the last six weeks of their pregnancies showed a clear preference for hearing this story over a similar story shortly after birth. Babies whose moms had not read *The Cat in the Hat* showed no such preference (DeCasper & Spence, 1986). If you're wondering how in the heck we know what infants prefer, one way to find out is to give them pacifiers that determine what they are exposed to. Even newborns will suck on a pacifier more readily when doing so leads to something pleasant than when doing so leads to something unpleasant (or less pleasant). Likewise Japanese quail who had been played specific selections of classical music *while they were incubating* preferred the specific musical selection to which they had been repeatedly exposed once they hatched (see Zajonc, 2001). There is even evidence that people come to prefer the specific foods their biological mothers consumed when they were in utero (or when they were breastfeeding; e.g., see Mennella, Jagnow, & Beauchamp, 2001).

As you may have already noticed, mere exposure effects travel well to different populations as well as to different situations. They are easy to demonstrate in the artificial confines of the lab as well as in more natural settings. They have been documented not only in people from cultures all over the world but also in animals as varied as Japanese quail, ravens, guppies, macaque monkeys, and rhesus monkeys. Consider a case study from Harlow's (1958) classic work with rhesus monkeys. You may recall that infant monkeys prefer soft and cuddly mothers who do *not* feed them over wire mothers who *do* feed them. But infant monkeys also appear to prefer *familiar* fake mothers over unfamiliar ones. By accident, one monkey in Harlow's studies was reared for its first six months by a surrogate mother *without a painted face*. When the researchers dutifully replaced the defective mother with a standard mother with a painted face, the young monkey "repeatedly screwed the new mother's head around so as to restore the beloved blank" (Brown, 1965, pp. 39–40). Apparently, a familiar face, even if it proves to be a blank one, is an important part of the magic of motherhood.

Putting all this together, we hope you can see that even a program of research based heavily on lab experiments can be high in external validity—at least it can if researchers are careful enough to conduct a wide range of experiments that examine all aspects of the OOPS! heuristic. Of course, Zajonc has not been the only researcher to document an experimental finding in a very wide range of ways. But doing so was certainly a signature of much of his thoughtful work. We should also note that we chose to use the mere exposure effect as an example precisely because the effect has proven to be very robust. There are some important effects in psychology that do not generalize as well as the mere exposure effect. For example, there is no evidence that ravens and monkeys experience cognitive dissonance in the unique ways in which people do.

But research on cognitive dissonance theory has gotten a great deal of attention in psychology because dissonance theory has survived a very wide range of critical experimental tests in people (e.g., see Aronson & Carlsmith, 1968). This has meant a wide range of studies that, taken together, have examined all aspects of the OOPS! heuristic.

So assuming that a researcher has uncovered a real effect, even seemingly artificial laboratory experiments can become a part of a research program that is high in external validity. But can passive observational studies ever be high in internal validity? As you already know, addressing a few very common confounds can take a passive observational study a long way in that direction. But what does this look like in practice? Let's take a look at a series of archival studies that, in our opinion, did an excellent job of maximizing internal validity. We hope you'll also see that these same archival studies had no shortage of external validity—especially if one views each unusual archival study as a real-world complement to a body of mostly experimental research that began in the lab.

GAUGING GAGES IN ARCHIVAL STUDIES OF SOCIAL COGNITION

False Consensus

One of the first researchers to step out of the lab to study social cognition in the real world was Brian Mullen (1983), who studied the **false consensus effect** (Ross, Greene, & House, 1977). This is the tendency for people to overestimate the percentage of others who share their beliefs or behaviors. Mullen believed this bias would still appear when avoiding it could help people win thousands of dollars in cash and prizes. Mullen also suspected (correctly) that the false consensus effect is larger for people whose attitudes or behavior place them in the statistical minority rather than the majority. To study the false consensus effect, Mullen capitalized on data from a TV game show (*Play the Percentages*). The key data points provided by game show participants were their estimates of the percentage of studio audience members who would be able to answer specific trivia questions (e.g., "What state did Hubert Humphrey represent in Congress?") Back when people still remembered Hubert Humphrey, 72 percent of audience members were able to answer correctly that Humphrey represented Minnesota.

Mullen observed clear evidence of the false consensus effect. Participants overestimated the percentage of others who knew the answers to questions when they *themselves* had known the answers to the questions. Second, as Mullen predicted, false consensus effects were larger than usual when people's own answers placed them in the statistical minority. The rare people who knew the answer to a difficult question were especially likely to overestimate the percentage of others who shared

their esoteric knowledge. And of course this was true even when people were trying very hard to guess correctly the percentage of audience members who did or didn't know something. As Mullen was quick to remind his readers, offering incorrect estimates of what the studio audience members knew in this gameshow cost many of these contestants dearly.

Mullen documented a false consensus effect with a slightly different operational definition than the one usually used in the lab, with a novel population, and in a very different situation than the lab, satisfying three of the four OOPS! criteria. But that's all about external validity. What about *internal* validity? It's pretty hard to imagine that a GAGES confound that would apply to Mullen's archival study without also applying to laboratory studies. For example, more highly educated participants *may* have known more of the answers to the trivia questions, but there is no reason to believe that being educated *in and of itself* would make people offer higher *consensus* estimates—or that this confound would happen in game shows but not in laboratories. Or consider geography. Certainly if a person happened to have grown up in Minnesota, that person would have been especially likely to answer the Hubert Humphrey question correctly. But this game show, like most other game shows, asked people trivia questions about all kinds of topics, which should have meant that geography was not a predictor of contestants' knowledge across all of the questions. Further, let's assume, for the sake of argument, that contestants from wealthy states like Maryland and Massachusetts had better knowledge of trivia than people from not-so-wealthy states like Georgia and Virginia (the home states of your two authors). Unless people in wealthy states were also socialized to believe that studio audiences are highly knowledgeable about gameshow trivia, this potential geographic (or socioeconomic) confound would not seem to be a problem. Gender, too, seems like an unlikely confound. Even if the questions were biased to be easier for men than for women, for example, there is no reason to think that men generally overestimate the percentage of others who know a lot about trivia. Further, if a critic of the study were somehow worried about a gender confound, it would be very easy to check to see if the robust false consensus effects Mullen observed were equally strong for women and for men.

Ethnic Stereotyping and Discrimination

A more sobering example of archival research in social cognition is Eberhardt, Davies, Purdie-Vaughns, and Johnson (2006) research on stereotypes and capital punishment. Eberhardt and colleagues identified criminal records from more than 600 men who had been convicted of murder in greater Philadelphia between 1979 and 1999. They then identified all of the cases ($n = 44$) in which a Black defendant had been convicted of killing a White victim. Based on previous work by Blair, Judd, and Fallman (2004), they suspected that Black men convicted of killing White victims would be more likely to receive a death sentence when they had a more stereotypically Black *appearance* than

when they did not. The researchers showed photographs of all of the selected Black defendants to students who knew nothing of the men's criminal status. These judges assessed "the stereotypicality of each Black defendant's appearance and were told they could use any number of features (e.g., lips, nose, hair texture, skin tone) to arrive at their judgments" (Eberhardt et al., 2006). Figure 3.4 shows two Black male volunteers who vary in stereotypicality.

One of the most methodologically impressive aspects of these archival findings is the fact that Eberhardt and colleagues controlled for six potential confounds known to be important predictors of sentencing decisions. These confounds included "(a) aggravating circumstances, (b) mitigating circumstances, (c) severity of the murder (as determined by blind ratings of the cases once purged of racial information), (d) the defendant's socioeconomic status, (e) the victim's socioeconomic status, and (f) the defendant's attractiveness." Further, their operational definitions of constructs a through e were based on well-established Pennsylvania statutes. Because the archival records did not include information on defendant physical attractiveness, the research team got blind raters to judge this. Even after controlling statistically for all six of these confounds, Eberhardt et al. (2006) found that Black men with a more stereotypically Black appearance were more likely than Black men with a less stereotypic appearance

Figure 3.4 Two Black men with *no criminal records* who vary in the stereotypicality of their appearance. The findings of Eberhardt and her colleagues (2006) suggest that if both men *were* to commit a crime, the man on the right would be judged more harshly.

Source: Hagiwara et al. (2012). Reprinted with permission from Elsevier.

to be given a death sentence. A follow-up study showed that when defendants had been convicted of killing *Black* rather than *White* victims, the stereotypicality of the men's appearance no longer made a difference for their sentences.

A consideration of GAGES reveals that this study controlled for many geographic confounds by staying near Philadelphia. Of course, OOPS dictates that it would have been even better to study more than one region of the United States. But no single study can do everything. The authors also appear to have controlled for the defendants' ages because age is often considered a mitigating factor in murder cases. They controlled for the gender of the targets by studying only men (though this raises the interesting question of what would happen in cases in which Black *women* had been convicted of murdering a White victim). The authors not only controlled for ethnicity but also deconstructed it (ethnic stereotypicality was an independent variable). They also controlled for the SES of both the defendants *and* the victims. The additional factors for which Eberhardt and colleagues controlled reveal that GAGES is not an exhaustive list. But the fact that these authors left no GAGES stone unturned attests to the importance of these variables as well as to the methodological sophistication of this sobering archival study.

Counterfactual Thinking and Emotions

Not all research in social cognition focuses on tragedies. Some of it focuses on triumphs. Medvec, Madey, and Gilovich (1995) studied athletic triumphs, including triumphs that don't always make people feel very good. Laboratory research on **counterfactual thinking** shows that when something good or bad happens, people often consider *counterfactual* (alternative) realities. Counterfactual thoughts sometimes create counterintuitive emotions. For example, missing a flight by two hours usually produces regret. But missing a flight by two minutes usually produces a lot more of it (Roese, 1997). When Medvec and colleagues conducted their archival studies of counterfactual thinking and emotions following real athletic performances, almost all previous studies had been conducted in the lab. Further, many of these studies were based on hypothetical scenarios ("How would you feel if . . . ?") rather than real outcomes. Medvec and colleagues put the factual back into the study of counterfactuals.

They did so by considering the emotional implications of earning gold, silver, and bronze medals in major athletic competitions. Most Olympic gold medalists must surely be on top of the world after their victories. At a bare minimum they end up on top of the medal stand, and their gold medals often bring them fame and fortune. By contrast, many silver medalists may feel the pain of knowing how close they came to winning. For bronze medalists, however, *two* things would have had to have gone differently for them to have won gold (e.g., both Usain *and* Justin would had to have pulled a hamstring). The most salient counterfactual for bronze medalists is probably

the fact that they could have easily finished in fourth place, earning no Olympic medal at all. This logic suggests that athletes might typically be happier with an inferior outcome (a bronze medal) than with a superior one (a silver medal).

To test this prediction, Medvec et al. (1995) recorded NBC's televised coverage of the 1992 Olympics. They then extracted every scene that showed a bronze or silver medalist (in any sport NBC chose to cover) the moment the athletes first learned they had finished second or third. They did the same thing for the period when athletes stood on the medal stand. Finally, they showed all of the video clips to a group of raters who were kept blind not only to Medvec et al.'s predictions but also to the athletes' order of finish. They also turned the volume to zero for all of the ratings so that raters would not be biased by the comments of any of the NBC sports analysts, especially Bob Costas. The raters simply judged each athlete's expressed happiness on a 10-point scale.

Medvec and colleagues found that, despite finishing third rather than second, Olympic bronze medalists looked happier than their slightly faster, stronger, and more coordinated peers. This was true both immediately after their performances and on the Olympic medal stand. Of course, these results alone do not say whether *counterfactual thinking* was responsible for the observed emotions. To address this, Medvec et al. performed a second set of archival analyses from the same Olympic TV coverage. This time they selected all of the available *interviews* with bronze and silver medalists and asked blind raters to judge the "extent to which the athletes seemed preoccupied with thoughts of how they did perform versus how they almost performed." This follow-up study suggested that bronze medalists were more focused on what they "at least" did whereas silver medalists were focused on what they "almost did." A replication study focusing on a state-level athletic competition confirmed this result. By the way, it would be interesting (and pretty easy) to reanalyze these data and code for where athletes were in their Olympic careers. Research on counterfactual thinking suggests that thinking about the future (especially how it could be different) can sometimes reduce the emotional consequences of counterfactual thinking (Strathman, Gleicher, Boninger, & Edwards, 1994). Based on this logic, we would expect the effects Medvec et al. documented to be larger than usual for athletes who were nearing the end of their athletic careers (e.g., because this could easily have been their last chance ever for a gold medal). For athletes competing in multiple events, one could also code for whether the athlete in question had any events *remaining* in this particular Olympic games. Seeing when an effect gets bigger versus smaller can often lead to important insights in archival research.

These archival results seem safe from any obvious GAGES confounds. It is not possible, for example, that men more often finish third than women. Moreover, this would only be a problem, even if it *were* true, if men were also chronically happier than women. Further, let's assume that people from Canada or Brazil smile a lot more than people from Uruguay or Belgium. Unless Canadians and Brazilians are also

especially prone to win bronze but not silver medals, it is very hard to explain these findings based on any kind of geographic confound. Moreover, if one were worried about this, it would be pretty easy to go back to these original data and code for the nationality of all of the silver and bronze medal recipients. One very real confound, however, is that in some Olympic events (e.g., wrestling, basketball), bronze medalists have just *won* a competition whereas silver medalists have just *lost* a competition. That's a real confound. In a supplemental analysis, Medvec et al. (1995) focused solely on events (e.g., track and field) in which there was no such confound. The bronze medalists still looked happier than the silver medalists. This archival research is also a standout when it comes to the OOPS! heuristic. It used novel operationalizations, it examined behavior in athletic events that took place on many different *occasions*, the *population* studied came from all over the globe, and the *situation* in which people were studied was radically different than the lab. In our view the authors of this study struck methodological gold.

SUMMARY

To put everything we have said in this summary chapter in just a few sentences, you should give some careful thought to both internal and external validity before you design and carry out any research project. Now that you know that experiments are an ideal way to maximize internal validity, for example, you might consider conducting an experiment on memory, helping, or athletic performance. The magic of random assignment should guarantee that you won't have to worry about any confounds involving individual differences. Further, now that you understand the OOPS! heuristic, you know that you should give some careful thought to whether (and if so how) your operational definitions of memory, helping, or athletic performance differ from those used in past research. And if you conduct any kind of passive observational study, you should be attentive to the five GAGES confounds. Finally, even if you are unable to conduct an experiment *and* unable to control for any GAGES variables, you now know to be very careful about your interpretation of your passive observational research findings. In short, we hope that if you are planning to conduct any research of your own in psychology, you are now in a pretty good position to do so in a way that maximizes the value (i.e., the internal and external validity) of your research.

STUDY QUESTIONS

1. Dr. Madd wants to see if people perceive more anger in ambiguous human faces if they themselves are currently feeling angry. In her past work, she has had a confederate (a trained actor who plays a role in an experiment) make insulting comments about the way real participants spoke and acted prior to getting them to rate the ambiguous

faces. Come up with at least one alternative operationalization of anger that would *not* require the use of a confederate—but which also would not make participants angry at the *experimenter* (which would be problematic).

2. Dr. White conducts research on college students at Iowa State University, where more than 75 percent of students are White. Further, the three largest ethnic minority groups in the United States (Latinos, Blacks, and Asians) collectively make up just under 10 percent of Iowa State students. Thinking about the OOPS! heuristic, what challenges might this pose for the external validity of Dr. White's research? What, if anything, can she do about it?

3. A survey researcher conducted a study with a large and representative sample showing that receiving high levels of social support is associated with better physical health. Further, this association held up even after controlling statistically for all five of the GAGES variables (e.g., even after controlling for the fact that highly educated people said they received more social support than less educated people). Based on what you know about the *problem of induction*, offer a reason why the researcher cannot conclude *for sure* that receiving a lot of social support causes people to be healthier.

4. A researcher who knew the director of a natural birthing facility interviewed 100 women who were planning to give birth at the center. Specifically, the researcher interviewed the mothers-to-be during their six-month check-ups at the birthing center. She asked the mothers-to-be if they planned to breastfeed their infants until the infants were at least one year of age. Imagine that exactly 90 percent of the 100 mothers-to-be who were interviewed answered that they planned to do so. Using the OOPS! heuristic as a guide, explain why the researcher *cannot* safely conclude that 90 percent of U.S. mothers today do, in fact, breastfeed their infants for the first year of life.

5. In a longitudinal study of infant and toddler toy preferences, a group of researchers at the University of Iowa examined how much time Iowa boys aged nine months versus 18 months spent playing with "gender-congruent" toys (such as cars and trucks) versus "gender-incongruent" toys (such as female dolls and stuffed animals). The researchers found that the nine-month-olds played with the gender-congruent toys about 60 percent of the time whereas the 18-month-old boys played with the gender-congruent toys about 90 percent of the time. This difference between groups was significant. Because this was a longitudinal study, the same group of boys served as participants when they were nine versus 18 months of age. First, identify the independent and dependent variables in this study. Second, explain how the longitudinal design addressed all of the concerns one might conceivably raise based on the GAGES heuristic.

6. Now critically analyze the external validity of this same longitudinal study. Using the OOPS! heuristic as a guide point out a few ways in which a follow-up study might increase the external validity of this basic finding.

7. Use a letter from the terms below (A–E) to label each study or finding that follows the list:

   A. archival research B. ethnography
   C. unobtrusive observation D. survey
   E. true experiment

___I. Use marriage records to show that grooms are, on average, about three years older than brides.

___II. In a study of persuasion, participants heard a message delivered at either 60 or 100 words per minute.

___III. Tina lived for two years in a foraging culture in Papua New Guinea—where she made systematic observations of gender, social status, and interruptions in daily conversations.

___ IV. Kelli took part in the 2010 U.S. Census reporting facts such as her age and marital status.

___ V. An archaeologist conducted genetic and microscopic analyses of food particles caught between the teeth of well-preserved Neanderthal skulls—to see how much meat Neanderthals ate.

8. Almost all methodologists argue that there is a big trade-off between internal and external validity. To some degree, we have repeated that argument here. However, if an experimenter conducted numerous experiments and truly followed all the advice offered via the OOPS! heuristic (e.g., using multiple operational definitions, studying a wide range of populations, etc.), how worried would she really have to be about the external validity of her entire set of laboratory findings? In other words, assuming there is a *de facto* ("in fact") trade-off between internal and external validity as research is often conducted, make an argument that (a) the trade-off can be reduced but never eliminated and (b) the trade-off can be eliminated completely. Don't forget to consider both passive observational and experimental research (and thus both GAGES and OOPS!).