CHAPTER 5

Experimental Design

Seven Lessons (Plus or Minus Two)

1. Avoid complex designs. You will be rewarded with happy participants (who provide better data), manageable analyses, and a focused test of your team's hypothesis.
2. Work as a team to design a strong experiment. Consider multiple perspectives, refrain from putting any single idea on a pedestal, and be willing to compromise.
3. Minimize unwanted effects by using random assignment and carefully attending to and controlling for extraneous variables that might interfere with your results.
4. Select your measures with care, because no matter how strong your independent variable may be, your team will not find an effect if the dependent variable is weak or inappropriate.
5. Evaluating whether a variable such as age or personality influences the strength of the relationship between your team's manipulation and the primary outcome is a search for *moderators*.
6. Think of any given experiment as the first in a series. Over time, multiple studies will come together to clarify the answer to a good research question.

I have so heavily emphasized the desirability of working with few variables and large sample sizes that some of my students have spread the rumor that my idea of the perfect study is one with 10,000 cases and no variables. They go too far.

—Jacob Cohen (1990, p. 1305)

Experimental design begins once your team has identified relevant theories and formulated predictions. Now it is time to transform your team's idea into something concrete. To launch our discussion of experimental design, let's begin with an example. We

would like you to spend the next few minutes writing a brief description of what excites you most about the process of research. But here is the catch: You are not allowed to use the letter *a* or the letter *n* in any of the words you write. Begin now. (The best researchers have a knack for simulating and therefore experiencing the phenomena they study. This means you have to try things out. So, we're serious: Start writing!)

Immediately after completing this writing exercise, quickly assess your current feelings. How energetic do you feel? *If you resemble most people who try to write without these two letters, you'll feel tired out.* (It took us a full two minutes to write the preceding, italicized sentence according to the rules, so we share this feeling.) This type of writing task requires a great deal of mental self-control, and after you have completed the task, the energy that fuels your self-control has been depleted. You might feel some combination of being tired and a bit frustrated, but at the same time engaged and perhaps a little relieved that you are done. Overall, your mood will not be particularly positive or negative, but you will feel like you have done some strenuous mental exercise. Indeed, exercising self-control is like exercising your muscles. It's draining.

So what? Here is where the story of self-control gets fascinating. Imagine you are reading this chapter in a bustling public setting such as a coffeehouse, the library, or a lounge (perhaps this is actually the case). Nearby, on the floor, you see a $5 bill. Would you pick it up? Would you then ask the people around you if they dropped the money? Or would you pocket it? Based on recent theory and research concerning self-control, right now you are *much* more likely to pocket the money than to ask others if they dropped it. Don't worry; we do not think you are a dishonest person. Instead, that annoying writing task sapped your internal reserve of self-control, and now you have less energy "left over" to do something else that involves self-control. It is much easier to give in to temptation and keep the money than it is to go through the effort of finding its rightful owner. Mead, Baumeister, Gino, Schweitzer, and Ariely (2009) designed two studies much like this that were published in the *Journal of Experimental Social Psychology*. The title of their article says it all: "Too Tired to Tell the Truth: Self-Control Resource Depletion and Dishonesty."

Before describing these studies in greater detail, we would like to remind you that there are a number of different approaches to carrying out research in psychology, ranging from descriptive (e.g., case studies) to correlational (e.g., opinion polls) to experimental (e.g., laboratory-based studies). Self-control could be studied with any of these approaches. A researcher could do a descriptive study by observing children diagnosed with attention-deficit/hyperactivity disorder in their classrooms in order to determine the situations in which their impulse control is most likely to break down (e.g., before lunch, after recess). A researcher could do a correlational study by surveying a random sample of adults and asking them to answer a questionnaire about a time when they acted impulsively. Or a researcher could do an experiment by bringing participants into the lab and randomly assigning them to experience one of two events (i.e., one designed to deplete their self-control resources and the other designed to maintain those resources) and then looking at the impacts of these two events on a subsequent measure of impulsivity. Table 5.1 provides examples of the most common research designs in psychology. For additional information about various research methodologies, we encourage you to consult a traditional research methodology text or a reputable online resource such as the Research Methods Knowledge Base (http://www.socialresearchmethods.net/kb/index.php).

TABLE 5.1 Three Types of Research Design in Psychology, With Representative Examples

Design type	Examples
Descriptive	Naturalistic observation Case studies Focus-group interviews
Correlational	Surveys Psychological testing Research using archival data
Experimental	Between-subjects designs, where participants are randomly assigned to either an experimental or a control group Within-subjects designs, where the same participant experiences more than one experimental condition

In this book, we focus exclusively on experimental design because experimentation is the single best approach for determining causal relationships, and we believe that mastery of experimental methods is an important goal for all students of psychology. We begin our discussion of experimental design by taking a look at Mead and colleagues' (2009) first experiment. Participants came to the research lab and were randomly assigned to write a short essay without using the letters *a* and *n* (the self-control *depletion condition*) or, alternatively, without using the letters *x* and *z* (the *nondepletion condition*). After participants completed this exercise, they were told that the experiment was over. (Yes, there is some deception going on here because the experiment was not over, and this would need to be justified in the researchers' IRB application.) Then, as part of a "separate" experiment, participants were given a sheet with 20 matrices of numbers, and their goal for each matrix was to find the two numbers summing to 10.00 (e.g., 7.79 and 2.21). They were told they would earn 25 cents for each matrix they solved. This search task was tricky, and participants had only five minutes to work.

Then came the key part of the experiment. Half of the participants were randomly assigned to have the experimenter score their worksheets and compensate them appropriately. This provided a baseline for participants' typical performance, and the average participant earned about 50 cents whether they wrote the draining essay (without *a* or *n*) or the easier essay (without *x* or *z*). This is important. Being drained of self-control does not seem to undermine just any intellectual task. Instead, it should undermine tasks that require *self-control.* Self-control is the ability to override, for instance, undesirable impulses. To test this, the experimenters had the other half of the participants score their own worksheets and take the appropriate amount of money out of an envelope containing twenty quarters. These participants were told to score their worksheets, recycle them using the room's paper shredder, pay themselves from an envelope containing $5 in quarters, and then tell the experimenter they were done. *Nobody would know if they cheated and awarded themselves a couple extra quarters.*

What did the experimenters find in the self-scoring condition? Clearly, participants were tempted to be dishonest. Even participants in the nondepletion condition cheated a bit. On

average, the nondepletion participants awarded themselves *one* extra quarter. But the real story is how participants who wrote the essay without the letters *a* and *n* responded to this tempting situation. By now you can anticipate the results. These participants had depleted their resources of self-control, and for them the impulse to take the easy money was much more difficult to override. Compared to their counterparts who earned only two quarters in the experimenter-scored condition, these participants awarded themselves *five* quarters on average. These participants didn't just fudge their scores a bit, they more than doubled their legitimate earnings. They were *stealing*.

Every good study has a moral—that is, a meaning that goes well beyond the specific research design. This is the hallmark of a well-conceived and well-designed experiment. In this study, the moral is that people generally resist the urge to be dishonest, but their success in doing so quickly deteriorates when they have access to fewer psychological resources enabling them to exercise self-control. If your self-control has been depleted in one domain, you are less likely to exert control in another domain. The implications of this relatively straightforward experiment should serve as a warning to all of us. Never go to an all-you-can eat restaurant that has wonderful desserts right after having a "delicate conversation" with your significant other. You won't eat healthfully. Never fill out a time sheet or prepare your tax return when you are trying to avoid thinking about missing the party that your neighbors are throwing next door. The temptation to tinker with your numbers is likely to be too great and could get you into serious trouble. And never leave a basket of candy out on Halloween expecting children to be honest and take just one piece each. The poor kids have already used up their self-control at previous houses as they put candy directly into their buckets, as told, rather than into their mouths.

DESIGNING YOUR STUDY

With the preceding examples in mind, you can begin the process of designing your team's study. At the outset, we should say that some people like this part of research because it focuses their efforts. Others feel like they are giving too much up when they narrow an otherwise big idea down to a specific experimental test. For just about everyone, crafting a new experimental design evokes a complex set of emotions. This is the moment when the ideas that inspire you, the theories that guide you, and the many practical considerations that constrain you (e.g., ethical requirements, limited time, the availability of participants) come together in one place. It is finally time to bring your team's research question to life.

Lesson 1: Avoid Complexity

We began this chapter with a tongue-in-cheek statement made by Jacob Cohen (1990), one of the most influential figures in statistics and experimental design in the behavioral sciences. The absurdity of his suggestion that "the perfect study is one with 10,000 cases and no variables" is meant to reinforce a key principle of designing studies: *Less is more.* Human behavior is breathtakingly complex, but experiments *must* be focused. So take it as a cardinal rule that you should manipulate and measure only a few select variables in your studies. This admonition may seem obvious. Simplicity is elegant. But researchers,

and especially students, too often propose overly complex research designs. After all, if you are interested in understanding a rich theoretical concept, and if your participants will be taking time out to come to the research lab or complete your online survey, why not expose them to a variety of experimental manipulations? This should enable you to test a number of hypotheses at once, right? And why not ask your participants hundreds of questions and have them complete every measure under the sun? This should enable you to detect and precisely specify each and every effect of the experimental manipulation, right? The answers to these four questions are, respectively, *don't do it, no, don't do it,* and *no!*

Complex designs typically backfire. There are many reasons for this, including participant fatigue, overly complex and cumbersome data, and the likelihood of diluting and therefore weakening your ability to test any particular hypothesis. In marketing research, phone surveys are kept to 8–12 minutes for a reason. Survey participants grow tired after a short while, and at that point their attention to the details of new instructions or questions wanes. As a result, participants' responses become less and less reliable. Have you ever taken an online survey only to catch yourself daydreaming or speeding up your responses halfway through? This is a typical reaction, and it is one that your team would like to avoid in your own studies. To the extent that an experiment is more engaging than a question-after-question survey, you can get away with a somewhat lengthier study. Likewise, participant fatigue has less of an effect on some phenomena, including many simple cognitive tasks. Still, you should streamline your team's study as much as possible.

Even if you can keep a person's attention for an extended period of time, having a large number of experimental manipulations or outcome measures can lead to overwhelming complexity when you try to analyze your data. We once ran three related studies in our lab, each of which included more than 100 loosely organized questions spread over two experimental sessions. One of these studies required participants to answer nearly 200 questions over four distinct stages of the study, including an initial experimental session and three time points during a lengthy follow-up session one week later. The data from these studies paralyzed one of us (who had won an award for teaching graduate-level statistics at Yale) for almost two years. Simply put, there were too many somewhat related yet somewhat distinct questions to manage and neatly reduce into a smaller number of compelling variables. And because of this, it was nearly impossible to tell a consistent story across the three studies. There are other problems associated with having too many variables (e.g., the increased likelihood of type I errors), but our point here is that you do not want to put your team in this position when it comes time to tackle the data.

Finally, complex studies with too many experimental conditions rarely provide clean, direct tests of any particular hypotheses. Moreover, they spread researchers' most precious commodity, participants, too thinly across conditions. So even if the theory behind your idea is complex, the ideal approach is to isolate and test pieces of the theory systematically, *one step at a time.* For example, you might hypothesize that a person's making a healthy snack choice depends on that individual's mood and the social setting. Great, go ahead and test this! But don't design a study that puts each participant into one of four mood states (e.g., happy, sad, afraid, neutral) and then asks the participant to choose a snack in one of three social settings (e.g., alone, in the presence of a friend, in the presence of a stranger). This design results in *twelve* separate experimental conditions (i.e., four mood conditions further divided into three social settings), and the number of interrelationships among

these conditions is potentially overwhelming. Your data will make a fool of you more often than not if you insist on running studies like this. Trust us, we've been there.

No matter how compelling your team's hypothesis, or how tempted you are to try out all possible variations of your question, you are better off conducting a series of much simpler studies that build from and inform one another. Your team might start off by comparing the influence of three mood states (e.g., happy, sad, neutral) in only one social setting, demonstrating that moods can, in fact, influence snack choice. You can then follow up with a study using the two mood states and two social settings that you think are most likely to interact. Does sadness result in unhealthy snacking when a person is alone but healthier snacking in the presence of a stranger, whereas a neutral mood leads to about the same level of somewhat healthy snacking regardless of the social setting? That would be a great finding. Then your team could build another study from there.

Lesson 2: Utilize the Design Team

How exactly does the research team come into play during the design phase? Unlike some parts of the research process, designing the study is not best accomplished through the divide-and-conquer strategy. Instead, the design process benefits tremendously from the absolute immersion of all team members. Entertaining multiple perspectives as you go about trying to develop your methods is an ideal place to start off. Imagine that your research team is interested in exploring the extent to which people mimic each other's facial expressions, posture, and movements without being consciously aware of doing so (a phenomenon known as the *chameleon effect*; Chartrand & Bargh, 1999). There are a number of ways to go about investigating this topic, ranging from observational studies (e.g., watching pairs interact at a coffeehouse) to experimental ones (e.g., seeing if participants can be led to mimic one another in the research laboratory). There are also individual differences that may be of interest to the research team, leading to other possible questions: Are women more likely than men to engage in mimicry? Are highly empathetic people more skilled at mimicry than less empathetic ones? Your team also might consider the effect of mimicry on the person being imitated. Under what conditions does the person become aware of being mimicked? Does being mimicked make an individual more or less fond of the person he or she is interacting with? Are therapists more effective when they mimic their clients?

As you and your teammates engage in the process of considering all possible designs, you will solidify a sense of togetherness and learn how to work best with one another. Everyone has something to contribute based on his or her unique perspective. This process is naturally chaotic at times, and it is sometimes frustrating. We liken the group process to a line dance, where you and your teammates are on one side of the dance floor and the design ideas you are considering are on the other. You should each take turns dancing with the ideas, joining together then moving apart, but always passing along each idea (no matter how attractive) to the person dancing next to you. Gradually the feelings of chaos will be replaced with confidence and control. And one of the secrets to gaining control is letting go. It is crucial to avoid putting any particular idea on a pedestal. Working together requires compromise, but compromise is not a painful process if you are willing to give up your own idea in order to entertain another.

Going back to the example of the chameleon effect, imagine that your team has decided to run an experimental study to investigate the extent to which a naive participant unconsciously imitates the behavior of a confederate in a laboratory-based personal interaction. At this point your team needs to *operationalize* the concepts that are of interest. By "operationalize," we are referring to the process of making the *independent variables* (i.e., what is manipulated in the experiment) concrete and the *dependent variables* (i.e., the outcomes) measurable. More specifically, what will the confederate do in this study? And what type of participant behavior will "count" as mimicry? The process of operationalizing your independent variables (IVs) and dependent variables (DVs) helps to clarify, crystallize, and sharpen the vision for your project. This is the moment when the team commits to a certain set of specialized materials in the hopes of finding a particular result.

It is through the process of putting together these materials that ownership is developed. In order to investigate the chameleon effect, you would consider questions such as the following: What aspects of mimicry interest us most (e.g., posture, facial expressions, body movements)? How closely scripted should the confederate's behavior and conversation be? What questions should we ask the participants before they meet the confederate? What should the *cover story* of our study be (that is, what should the participants think is happening in the study so that they won't guess the purpose of the research)? What questions should we ask the participants after their interactions with the confederate? And how in the world are we going to keep track of and quantify our primary dependent variable, imitation of the confederate? As you continue to conduct group brainstorming about the design, feel free to let the chaos enter and leave the conversation. Play around with ideas and engage in freewheeling discussion until you feel comfortable with the methodology. At a certain point the number of unanswered questions will seem overwhelming, but over time the answers will outnumber the questions, and ultimately you will have a neat package of methods and a tentative plan.

TRANSLATING YOUR PLAN INTO AN ELEGANT METHODOLOGY

Let us remind you again that simplicity of design is crucial. There clearly are a number of directions that one could take any given experimental question, and the reality is that most published research today describes more than just a single study. It is common, and in fact expected, for researchers to collect and report data on two or three studies before they are able to "package" a compelling set of studies for publication.

The most basic experimental designs set up a comparison between a treatment condition and a control condition (or conditions), in which participants are not exposed to the "active ingredient" of the treatment condition. In the study we described at the start of this chapter on the depletion of self-control, participants in the "experimenter-scored" conditions were considered to be controls because they were not given the opportunity to cheat on the task. Their data were useful in providing a baseline for the number of correct responses participants produced under conditions of depleted versus nondepleted self-control. But the primary hypothesis that depletion of self-control would lead to cheating was not tested directly among these control participants. The "treatment" or experimental condition in this study was the self-scored group. These participants also were randomly

assigned to have their self-control resources either depleted or not, but unlike participants in the control condition, they were given the opportunity to cheat. Moreover, within this group, participants who did not have their self-control resources depleted by the more challenging writing task served as controls for the focal "treatment" group: participants whose self-control was depleted and were given the opportunity to cheat.

Experimental Control

To shape the remainder of our discussion of concepts that are crucial during the design stage, we will take Chartrand and Bargh's (1999) series of research studies on the chameleon effect as an example. Chartrand and Bargh conducted three studies, each building on the last, in order to create a more comprehensive picture of the natural human tendency toward imitation. Let's begin with Experiment 1. The hypothesis for this first study was that participants would unintentionally imitate the facial expressions and bodily movements of a confederate, who was a member of the research team but was presented to each true participant as another student taking part in the study. Each participant worked with the confederate on a task that was ostensibly part of a process of pretesting experimental measures. The participant and confederate took turns describing pictures that had been taken from magazines. After interacting with the first confederate, the participant "switched partners" and worked with another confederate. Depending on condition, the confederate in the first interaction smiled and rubbed his or her face or, alternatively, avoided smiling and shook his or her foot. The second confederate always exhibited the opposite set of behaviors from the first confederate (e.g., if confederate 1 smiled and rubbed her face, then confederate 2 avoided smiling and shook her foot). The confederates' behavior was the independent variable in this study. All participants were videotaped, and their behavior was later coded for smiles, face rubbing, and foot shaking. The participants' behavior as recorded by the coders was the dependent variable in this study.

In any research study, it is crucial to attend to *experimental control,* or the process of protecting the integrity of the experiment's conditions. Researchers accomplish this by minimizing and/or measuring any variables that might have an unintended influence on participants' responses to the experimental manipulations. In an investigation such as this one concerning the chameleon effect, the question to ask yourself is this: What variables other than the experimental manipulation might interfere with the observed results? One concern that Chartrand and Bargh had was that some people might be more smiley in general or they may be chronic foot shakers or face rubbers (no, these aren't the official scientific terms for such behaviors). What could you do to determine whether this is true? It would not be a good idea to ask participants, "How much do you smile?" in the prescreening questionnaire. Not only are people notoriously bad at giving objective responses to questions like that, but also such a question might give away the purpose of the study. Instead, Chartrand and Bargh videotaped participants for one minute prior to their interactions with the confederate in order to gather "baseline" observations of the frequency of their smiling, foot shaking, and face rubbing. These observations were coded and accounted for in the data analysis.

Another way in which experimental control was introduced into this paradigm was by having the participant interact with two different confederates. In this way, the participant

served as his or her own control; the experimenters were able to see if they could produce one set of behaviors (e.g., face rubbing and smiling) during the first part of the study and another (e.g., foot shaking and not smiling) in the second part of the study. Being able to demonstrate one effect and then changing the nature of the effect within the same experimental session can be a very powerful application of experimental control. In addition, the researchers had the legitimate concern that participants might imitate one research assistant (i.e., confederate) more than another. If so, it would be problematic to have participants interact with just one confederate, because any observed imitation could be a product of the confederate in particular rather than the chameleon effect in general.

Randomization and Counterbalancing

When we first began to describe the methods of this experiment, you probably noted that we mentioned that participants were *randomly assigned* to experimental condition. This means that there was nothing systematic in how participants were dispersed across conditions of the experiment (i.e., they were assigned by chance). Remember, the second confederate always performed the behaviors that were *not* demonstrated by the first confederate, and all participants interacted with both of the confederates. The confederates demonstrated two of four behaviors (smiling or not; foot or face movement), and the likelihood that the participant was assigned to one particular set of behaviors in the first interaction also was determined by chance. The experimenters made sure to *counterbalance* the confederates' behaviors, so that confederate 1 performed each mannerism and facial expression as often as confederate 2. In counterbalancing, the researchers make sure that all possible orders in which participants are exposed to elements of the experiment are utilized. Counterbalancing is deemed important whenever there is a concern that the order in which a person experiences aspects of the experiment could matter, because of familiarity, practice, or fatigue. Attention to random assignment and counterbalancing is crucial because researchers need to rule out the possibility that something other than the experimental manipulation is responsible for the study's observed effects. For instance, what if all participants who completed the study during the month of March were exposed to one sequence of events, whereas all participants who completed the study in April were exposed to another sequence? There might be additional systematic differences between conditions (e.g., the weather during these months) that could account for the results. Similarly, what if the same confederate were always the nonsmiling foot shaker? It's possible that incidental characteristics of the confederates (e.g., differences in attractiveness) could account for the results. Random assignment and counterbalancing help to address these concerns.

SELECTING MEASURES

To maximize the likelihood of finding an effect, most researchers emphasize conducting studies with an adequate number of participants. Two other considerations are equally important, however. Ideally, your team should increase the strength of the manipulation as much as possible without making the hypothesis so obvious to participants that they simply try to confirm your predictions (i.e., fall prey to demand characteristics).

Moreover, your team should strive to increase the precision of your dependent variables (i.e., decrease the statistical noise or error in your measures). This, along with choosing appropriate measures for your experiment, is a crucial part of the design process.

Measures can be placed at different points within a study. Many studies include "pre-questionnaires" to assess variables that will be taken into consideration during data analysis. This is especially useful in evaluating *changes* that occur as a result of the experimental manipulation. Of course, the questions that you ask after the experimental manipulation are your *primary dependent variables*. However, many researchers place demographic questions that are not likely to be affected by the experimental paradigm (e.g., self-reports of age and sex) at the very end of the experiment, immediately prior to debriefing. This helps to prevent participants from becoming fatigued prior to the experimental manipulation and primary measures. In some cases, these demographic data might relate directly to your hypotheses (e.g., women might be more likely than men to demonstrate the chameleon effect). This is a question of moderation, which we discuss at the end of this chapter. In other cases, demographic data might enable you to identify and eliminate the influence of nuisance variables on your primary measures (e.g., younger people might be much more likely than older people to shake their feet).

Coding Participant Behavior

Chartrand and Bargh obtained the data for their experiment by "coding" videotapes of the participants. Participants were told as part of the informed consent procedure that their responses during the experiment would be videotaped. Then, after all of the participants had completed the study, two "judges" who had not been involved in the earlier part of the experiment viewed and rated the frequency of each participant's facial expressions and mannerisms. These judges were *blind to experimental condition,* which means they did not know what behaviors were being exhibited by the confederate. The researchers achieved this by positioning the camera so that only the participant was visible on the video recording. Three distinct time periods were coded: the one-minute baseline, the interactions with confederate 1, and the interactions with confederate 2. The coding resulted in ratings of the number of times each participant smiled, shook his or her foot, and rubbed his or her face.

Lest you get the impression that the coding of such videotapes is as straightforward as a simple behavioral count, let's take a moment to do a reality check. What were some of the stumbling blocks that these researchers faced when it came time to translate the video material into usable data? The overarching issue that challenges every study that involves coding of data is *interrater reliability,* or the extent to which different judges evaluate a particular phenomenon in the same way. Adequate interrater reliability is notoriously difficult to achieve, especially if there is any ambiguity in what should be coded, which causes judges to guess when they rate certain things. In the evaluation of the chameleon effect, interrater reliability was lower for face rubbing than it was for the foot shaking or smiling. There are a number of reasons participants might rub their faces (e.g., to move their hair away, to scratch an itch), and some of these instances might be irrelevant to "mimicking." The best that the researchers could do was to create a detailed coding sheet (which included a description of the specific nature of the participant's face rubbing) to increase the reliability of the ratings across judges.

Another question about measurement is also relevant here: Is the *number of times* the participant engaged in a behavior actually the best measure of mimicking? Or would it be more informative to rate the *length of time* the participant smiled, shook a foot, or rubbed his or her face? As it turns out, the experimenters coded both variables, but because the two sets of ratings were so similar (that is, they were highly correlated), only the *number of times* ratings were discussed in the results.

This study highlights just one example of the types of measures experimenters can use. Although using video recordings of participants is appealing for many reasons (e.g., the measure gets at "real" behavior), coding the material reliably can be quite challenging. Not only do you have to worry about interrater reliability, but also you have the challenges associated with obtaining the necessary physical space and equipment to conduct this type of study. Unless your team is working with a research adviser who has recording equipment in his or her lab, it is unlikely that you will be able to create the right environment in which to collect this type of data. Thus, your team must consider using other types of measures.

Utilizing Existing Measures

Potential dependent variables come in many forms. We have just described how researchers can utilize observations (e.g., behavioral coding) as a study's primary measures. However, self-report measures are the most common type of measure used in psychological research. Self-reports come in many forms, ranging from simple reflections by participants on their own thoughts or feelings (e.g., "liking of partner") to multi-item scales that have been developed and validated by other researchers (e.g., the Young Adult Alcohol Problems Screening Test developed by Hurlburt & Sher in 1992).

The selection of appropriate measures is crucial, because no matter how strong your experimental manipulation is, you will not be able to find an effect if your dependent variables are ambiguous, irrelevant, or otherwise flawed. We encourage you to rely on the strengths and expertise of your team members throughout the process of choosing measures. For example, perhaps your team will be bringing romantic partners to the lab in order to study the impact of relationship satisfaction on expressions of emotion during a problem-solving task. Your hypothesis is that greater relationship satisfaction will be associated with a balance between positive and negative emotional displays, whereas lower relationship satisfaction will be associated with predominantly negative *or* predominantly positive (that is, unbalanced) emotional displays. You decide to use the Dyadic Adjustment Scale (Spanier, 1976) as your measure of relationship satisfaction (the IV). But how will you measure "expressions of emotion" (the DV)?

Again, being part of a team is a great advantage here. We suggest having each team member search out ideas, keeping in mind that there are a number of types of measurement techniques (e.g., observational, self-report, psychophysiological) that you might use. One team member might return with a well-established and validated self-report mood questionnaire such as the Profile of Mood States (Cella et al., 1987), which is a brief measure of transient mood. Perhaps another team member is a research assistant with a professor who studies emotion. As a result, she has been trained in the Facial Action Coding System (Ekman & Friesen, 1978), a method of coding emotion-specific facial movements

(a system popularized, at least loosely, in the 2009–2011 television drama series *Lie to Me*). She therefore proposes that the participant dyads' interactions should be videotaped so she can code their moods. All smiles are not the same, you learn, and she can distinguish an insincere and voluntary "Pan American smile" from a sincere and involuntary "Duchenne smile" (the key is the additional contraction of the inferior part of orbicularis oculi, in case you were curious). This team member argues that taking videos of participants and later coding participants' facial expressions would be worthwhile because self-reported emotions could be biased.

As for you? You worry less about the biases associated with self-reports and more about a questionnaire relying on participants' faulty memories of their emotions during the interaction. You would prefer a "real-time" (presumably more accurate) measure of mood. The Ekman coding system would address this concern, but you are wary of the time and effort involved in the painstaking process of coding participants' each and every facial expression. As an alternative, you propose having participants use "mood dials" throughout their problem-solving session. The mood dial is a relatively simple, handheld device that is connected to a computer via a USB port. The dial is 360 degrees and color coded from dark blue (very unhappy) to bright orange (very happy). Participants would be told to start their conversation with their dials set to the midpoint (neutral) and then turn them toward blue as their feelings become more negative and toward orange as their feelings become more positive. Not only does this method capture data on a person's mood as it happens, but in your study it would allow you to map each partner's data onto those of the other, so that mood-related questions (such as "To what extent does relationship satisfaction predict congruence in the mood of partners?") could be answered in real time. By sharing with one another the experiences and knowledge that you possess individually, you and your teammates can together shape the design of your experiment in a way that makes use of the best possible measures.

Creating Your Own Measures

In some cases it may be appropriate for your team to create your own measures. Perhaps, for example, you are researching how a college student's place of residence (IV) affects his or her perception of the college experience as a whole (DV). Although other researchers surely have created scales to assess students' perceptions of their college experiences, it may be important for you to tailor questions so that they include statements specific to your own campus. For example, you may want to assess the extent to which students see the college as (a) a community of scholars, (b) committed to diversity, and (c) cultivating civic engagement. It isn't likely that any preexisting, validated scales tap into these three constructs in a way that would be meaningful to your participants, so you will need to create a measure.

Creating a measure is not an easy or quick process, and the best measures are thoroughly pretested before they are used within a larger study. How does one go about doing this? First, some definitional issues. A *construct* is a specific psychological attitude or property that you, as the researcher, are attempting to measure. When you *operationalize* your dependent variables, you first will need to identify how many distinct constructs you plan to measure. (The student perceptions listed above might map

neatly onto three constructs; alternatively, there might be more than one distinct aspect of, say, "cultivating civil engagement" that should be measured as separate constructs.) After your team has identified the constructs, it is time to come up with particular questions that tap into each of them. Our advice is to develop three to four reliable questions for each construct. The questions will be worded slightly differently and susceptible to somewhat different biases, but combined together they will help your team "triangulate" on the overarching construct.

As an example, take the idea of assessing the extent to which students on your campus feel as though they are immersed in a *community of scholars.* There are a number of possible statements you could ask your participants to endorse (using a scale from 1 = *strongly disagree* to 7 = *strongly agree*): "I feel engaged intellectually both inside and outside of the classroom"; "I rarely have conversations with friends that draw on ideas that I learned in my classes"; "I am part of a campus culture that values lifelong learning." Each of these questions taps into the *community of scholars* construct. As you probably noticed, the second item is worded in the direction opposite that of the first and third, so that a high score (7) on that item reflects a belief that the campus does not cultivate a community of scholars. We encourage your team to include such *reverse-coded* items in your scales for one simple reason: Sometimes participants answer questions somewhat mindlessly, giving very similar, often positive ratings across all items. By having some items worded in the opposite direction from others, you will increase participants' attention to the questionnaire, and the combined measure will correct somewhat for any positive-response bias. Ultimately, the hope is to increase the overall *validity,* or accuracy, of participants' responses.

Earlier in this chapter we discussed interrater reliability, but when it comes to survey design, a different type of reliability matters: *internal consistency reliability*. This term refers to the idea that a set of items designed to measure a particular construct should be consistent with each other in a given survey. Reliability estimates range from 0 (no consistency whatsoever) to 1 (complete overlap). A number of measures are available for estimating internal consistency reliability, but the one you are likely to encounter most frequently (particularly when there are a large number of items to assess) is Cronbach's alpha or α (Cronbach, 1951). A discussion of how to calculate α is beyond the scope of this chapter, but you can find a comprehensive description of this and other measures of reliability online at the Research Methods Knowledge Base website (http://www.socialresearchmethods.net/kb/reltypes.php). In Table 5.2 we provide a list of these and other considerations that you should be mindful of when creating measures.

An Application of Measurement Concepts

The overarching principle in experimental design is to measure specific, predicted outcomes while minimizing all influences on the participant except for those that are deliberately introduced by the experimental paradigm. Going back to Chartrand and Bargh's line of research, how might the chameleon effect be measured using self-report questionnaires? Could you simply ask participants to chart how often they match their own physical movements to those of other people? Most likely not. It is doubtful that people have an unbiased view of the extent to which they mirror others; indeed, the definition of the

TABLE 5.2 Twenty Helpful Tips to Consider When Designing Measures

1. Give your team enough time for the initial brainstorming and revision process.
2. Identify the independent variables you intend to manipulate in your study.
3. Think about the broader, theoretical concepts motivating each measure.
4. Brainstorm questions—the more the better (at least initially).
5. Make sure your questions cover all aspects of your predictions.
6. Meet with your research mentor and teammates to compare ideas for measures.
7. Whenever possible, adapt constructs from previous research, especially if your team hopes to replicate past findings.
8. Select three to four items for each subscale (i.e., construct).
9. Pay attention to internal consistency reliability when measuring constructs.
10. When using established measures, locate an original article that reports reliability statistics and a factor analysis of the items, which can guide the selection of a subset of items for use in your team's experiment.
11. As a general rule, avoid yes/no questions because they reduce the statistical power of any test. Instead, use scaled items such as Likert-style 7-point scales or semantic differentials.
12. Reverse the direction of some questions in order to correct for response biases.
13. Ask participants demographic questions in order to test hypotheses related to potential moderators, or to identify and eliminate the influence of nuisance variables on your primary dependent variables.
14. Edit your questions for content, word choice, and tense.
15. When necessary, edit the phrasing of questions to fit the end points of the scale.
16. Rearrange the ordering of the questions to enhance the overall flow.
17. Confirm that the questions are phrased in a way that does not bias or distort participant responses.
18. Keep in mind Cohen's admonition to minimize the number of DVs and delete unnecessary items.
19. Have all team members and your research mentor scrutinize your measures before your team pilots the experiment.
20. Be prepared to do additional editing and rewriting of items based on feedback from pilot participants.

chameleon effect includes the idea that it is an automatic, passive, nonconscious process. So we remind you again of the importance of utilizing your whole team in designing your study. Some of the constraints that you and your teammates must impose on yourselves are practical ones. In this case, you should reach a point where you realistically assess the resources your team can access for the purpose of your study. If you are unable to investigate the participant's likelihood of *demonstrating* the chameleon effect, how about investigating the impact of the chameleon effect *on* the participant? Would the participant report greater comfort, liking, and connection with a person who mirrors his or her behavior? This might be a more testable hypothesis if you plan to use self-report measures.

Chartrand and Bargh asked this very question in their second experiment. In this study, the participant again interacted with another "participant" (actually a confederate), taking turns describing what they observed in a series of photographs. The experimental design was relatively straightforward. Half of the participants were randomly assigned to a confederate who was instructed to imitate the behavioral mannerisms of the participant; other

participants were assigned to a confederate who was instructed to engage in neutral, unremarkable mannerisms. The primary dependent variables were participants' self-reports of how much they liked the confederate and how well they felt the interaction went. Here is the researchers' description of the measures, as stated in the methods section of their write-up of the experiment:

> The key items read, "How likable was the other participant?" and "How smoothly would you say your interaction went with the other participant?" To help camouflage the hypothesis of the study, we embedded these two items among eight other questions that asked about the task itself and the group format (e.g., how easy or difficult it was for them to generate responses to the photos, and whether they thought the various photographs went well together as a single "set"). All items were rated on 9-point scales (for the smoothness item, 1 = extremely awkward, 9 = extremely smooth; for the likability item, 1 = extremely dislikable, 9 = extremely likable). (Chartrand & Bargh, 1999, p. 902)

We would like to believe that Cohen (whom we quoted at the outset of this chapter) would have approved of Experiment 2. This study included one independent variable (mirroring or not mirroring the participant) and two primary dependent variables (liking and smoothness of interaction). This makes for a simple and elegant test of the impact of the chameleon effect on participants.

Before we transition to our next topic concerning experimental design, it only seems fair that we should report the results of Chartrand and Bargh's Experiments 1 and 2. What were the hypotheses? And were they supported? In the first study, the researchers hypothesized that the participants would imitate the facial expressions of the confederates and that this process of imitation would be a nonconscious experience. The results supported these hypotheses. Indeed, participants smiled more times per minute with a smiling confederate than they did with the nonsmiling confederate. In addition, the predicted relationship between confederate behavior (foot shaking, face rubbing) and participant behavior was observed. These results are depicted graphically in Figure 5.1. Finally, at the end of the experimental session, participants were asked if anything stood out to them about the confederate's behavior, including mannerisms or way of speaking. Out of 35 participants, only 2 stated that they noticed something unusual: One noted that the confederate made hand motions while speaking, and a second noted that the confederate slouched. In short, none of the participants noticed the expressions or mannerisms being studied.

In the second study, Chartrand and Bargh hypothesized that the chameleon effect serves to increase liking and feelings of comfort during interpersonal interactions. As predicted, participants in the experimental condition reported liking the confederate more and feeling that the interaction went more smoothly in comparison with those in the control condition. Further, just 1 of the 37 participants in this study reported noticing anything unusual about the confederate's behavior during the interaction, again suggesting that effects of mimicry come about through a nonconscious process. Taken together, the two experiments support the idea that people do engage in automatic, nonconscious mimicry of their interaction partners and that there are social benefits (e.g., increased liking) that accompany this phenomenon.

FIGURE 5.1 Results of Chartrand and Bargh's (1999) study showing a strong correspondence between confederate and participant behavior.

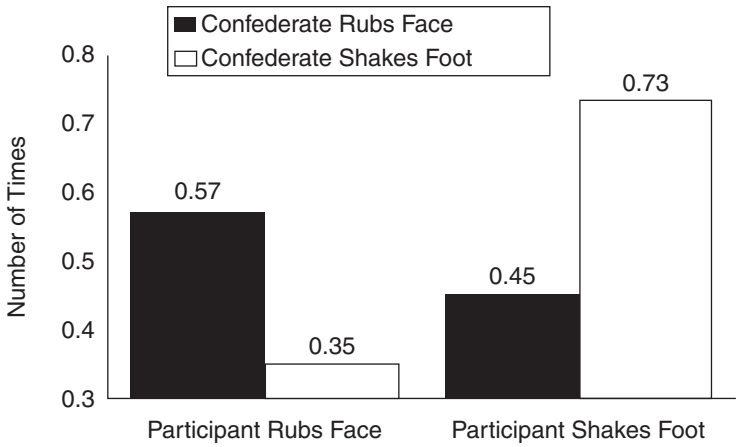

Figure 1. Number of times participants rubbed their face and shook their foot per minute when with a confederate who was rubbing his or her face and a confederate who was shaking his or her foot.

INVESTIGATING MODERATORS

In this chapter we have focused quite a bit on the chameleon effect. What questions about this effect remain? The answer is likely to be "many," but we will highlight just one more: Is it possible that some people are "better" chameleons than others? If so, what distinguishes skilled imitators from less skilled ones? Chartrand and Bargh wondered if any personality variables might predict how likely a person is to engage in mimicry. When researchers are interested in determining the extent to which a third variable (e.g., personality) influences the strength of the relationship between the independent variable (e.g., the confederate's body movements) and the dependent variable (e.g., mimicry by the participant), they are exploring a concept called *moderation*. In other words, the researchers are investigating whether participants' personality characteristics *interact* with the confederate's behavior in predicting the outcome (participant mimicry). If such an interaction exists, participants' personality characteristics are said to *moderate* the extent to which participants demonstrate the chameleon effect.

If your team is interested in looking for potential moderators, it is crucial to go back to a process of brainstorming. "Personality," for example, is a broad concept, and greater specificity is needed if you are to develop a meaningful hypothesis about how personality might moderate an effect. In their third (and final) experiment, Chartrand and Bargh hypothesized that a person's tendency to be empathetic (which is an individual difference characteristic) would predict that person's likelihood of mimicking the confederate. How did they measure empathy? A number of well-tested questionnaires have been developed

to measure individual differences in empathy, and if your experiment includes a concept such as this, it is critical that you avoid reinventing the wheel. This is an instance in which a wise research team goes back to the literature, investigates how others in the field are studying the concept, and brings back the best current thinking and practices to inform the team's own investigation.

What Chartrand and Bargh discovered was that empathy is not necessarily a single concept. Instead, theorists tend to distinguish the emotional part of empathy (e.g., feeling another person's feelings) from the cognitive part of empathy (e.g., being able to take another person's point of view). Which component of empathy, the emotional or cognitive one, might moderate the chameleon effect? Chartrand and Bargh decided to set up a horse race between the two competing moderators. That is, they measured both components of empathy in order to see the relative impact of each one. And, of course, they went into their third experiment with a hypothesis: They predicted that the cognitive component of empathy would be a more important moderator of the chameleon effect than the emotional component. They made this prediction because their earlier studies suggested that the chameleon effect can occur even in the absence of an emotional response or connection to another person. (If you recall, Experiments 1 and 2 involved a picture-rating task that did not elicit much if any emotional connection between the participant and the confederate.) As a consequence, the researchers predicted that participants who pay more attention to others and are skilled at taking another person's perspective (that is, those who possess the cognitive component of empathy) would be more likely to display the chameleon effect than those who do not have such a tendency. Furthermore, they predicted that this cognitive component of empathy alone, *not* the emotional component, would moderate the effect of the confederate's behavior on the participant.

The procedure for Experiment 3 was similar to that used in the first two studies. The participant and confederate completed a task in which they judged pictures, and this time the confederate engaged in face rubbing *and* foot shaking throughout the entire interaction. The confederate's facial expression remained neutral. As in Experiment 1, the participants were videotaped, and their mannerisms were coded for the amount of time they engaged in each type of movement. After the interaction with the confederate, participants completed the empathy questionnaire, which measured both the cognitive and the emotional aspects of empathy.

But why did the empathy questionnaire come *after* rather than *before* the interaction with the confederate? Does the order of manipulation and empathy measure matter? The concern here is that one aspect of the study (interacting with the confederate) might influence another aspect of the study (self-reported empathy) or vice versa. Recall that the prediction is *not* that confederate behavior should influence participant empathy. Instead, the prediction is that participant empathy will strengthen or weaken the extent to which mimicry occurs. This means that the researchers do not want responses on the survey to be affecting *or* affected by the interaction with the confederate. One way of handling this concern is to separate measurement of the moderator from the experimental paradigm. In this case, the researchers could have administered the empathy questionnaire days or weeks in advance of the experiment. As you might imagine, however, this can be difficult to achieve. The reality of experimental work is that resources (e.g., participants) are limited, and it might be difficult to count on having access to the same participants over time.

If practical limitations require you to measure a potential moderator within the experimental paradigm, as was the case in Chartrand and Bargh's study, you ultimately must ask which concern is greater: the measure's impact on the experimental manipulation or the impact of the experimental manipulation on the way participants respond to the measure. The answer to this dilemma is not an easy one, but in this case there was something very important to preserve: the believability of the cover story. If the participant first fills out a questionnaire that asks about his or her ability to take another person's perspective and the extent to which he or she feels an emotional connection to another person's experiences, the participant might be *primed* to think about these aspects of empathy during the experimental interaction. Moreover, the participant might become suspicious of the study's real intent and begin to question the behavior of the confederate. This would present great problems for the integrity of the research design. In contrast, empathy is considered to be an individual difference characteristic that is stable over time. So although a participant's responses to the empathy questionnaire conceivably might be influenced by the interaction he or she just had with the confederate, the influence is likely to be very small. It is through weighing these pros and cons that researchers determine the order of manipulations and measures in an experimental design.

So what did Chartrand and Bargh find? As predicted, they found that participants who scored higher on the cognitive component of empathy (perspective taking) rubbed their faces and shook their feet significantly more times per minute than did those who scored lower on this measure. However, the emotional component of empathy (empathic concern) did *not* moderate the chameleon effect. In other words, only the participant's perspective-taking ability (and not his or her empathic concern) influenced the extent to which the participant imitated the confederate. Again, the researchers' hypotheses were supported, and their findings added to the overall understanding of the chameleon effect.

Conclusion

As you have likely observed in reading this chapter, there are numerous considerations to juggle during the experimental design process, and we have covered only a subset of them. The overarching challenge your team must meet in designing your study is to be highly organized, systematic, and meticulous. You will find that the beginning of the design process is characterized by chaos; your team is likely to consider manipulating and measuring everything and anything. Once you target a few key variables, the chaos will subside and a sense of order will begin to take over. From this point on, feasibility should be a key component of all your design decisions; there are clear restrictions on what can be done in a single study (ethical constraints, time limitations, and limited finances being just a few).

When you consider concepts such as validity (i.e., is your experiment measuring what you think it is?), we encourage you to focus on *internal validity* rather than *external validity*. That is, it is critical to strive for an elegant, tightly controlled, clearly orchestrated study, regardless of whether the study takes place in the lab, online, or on paper. It is less important to be able to conclude, at the end of just one study, that your findings have broad implications for the "real world." With any line of research, the experimenters must aim to understand the effects as completely as possible, but this understanding does not

come after one or two or even three studies. Over time, a team of researchers will work to understand multiple facets of a phenomenon. Scientists use the phrase *line of research* deliberately. Each individual experiment continues along this "line," moving the question forward in a systematic fashion. We challenge your team to let go of the *ideal* design in order to pursue the best *possible* design, one small step at a time.

References

Cella, D. F., Jacobson, P. B., Orav, E. J., Holland, J. C., Silberfarb, P. M., & Rafia, S. (1987). A brief POMS measure of distress for cancer patients. *Journal of Chronic Diseases, 40,* 393–342.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76,* 893–910.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Ekman, P., & Friesen, P. (1978). *Facial Action Coding System: A technique for the measurement of facial movement.* Palo Alto, CA: Consulting Psychologists Press.

Hurlburt, S. C., & Sher, K. J. (1992). Assessing alcohol problems in college students. *College Health, 41,* 49–58.

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., & Ariely, D. (2009). Too tired to tell the truth: Self-control resource depletion and dishonesty. *Journal of Experimental Social Psychology, 45,* 594–597.

Spanier, G. B. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family, 38,* 15–28.